

PREDICTING SELECTIVITY AND THE
FUNCTIONAL IMPACT OF CLINICAL CANCER
MUTATIONS FOR SMALL MOLECULE KINASE
INHIBITORS USING PHYSICAL MODELING

by

Steven K. Albanese

A Dissertation

Presented to the Faculty of the Louis V. Gerstner Jr.,
Graduate School of Biomedical Sciences,
Memorial Sloan Kettering Cancer Center
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

New York, NY

February 2019

John D. Chodera, PhD
Dissertation Mentor

Date

© 2019 Steven K. Albanese

ALL RIGHTS RESERVED

ABSTRACT

Small molecule kinase inhibitors have become a major focus of drug development for treating cancer, which accounted for 610,000 deaths and 1.7 million diagnoses in the United States in 2018 alone. Currently, there are 44 FDA approved small-molecule kinase inhibitors. The dominant paradigm for designing such inhibitors has been to optimize maximally selective ligands for a single target. Unfortunately, many such inhibitors fail in clinical trials due to a lack of efficacy and clinical safety. Tumors can evade inhibitors through multiple routes of resistance, including upregulation of a second kinase, mutations in the target kinase, or amplification of the target kinase. On the other hand, toxicity arises from on-target inhibition of the wild type kinase or off-target effects of promiscuous small molecules or their metabolites. ATP-competitive kinase inhibitors have great potential for promiscuity, as there are over 520 kinases in the human kinome that each bind a common substrate, ATP. Further, advances in sequencing technology have enabled the generation of datasets of disease associated alterations rich in missense mutations in kinases. While this technology has been particularly transformative in the field of oncology, where many patients are treated with kinase targeted therapies, most kinase missense mutations are rare, making it difficult to assess their functional impact. Physical modeling can provide a route for predicting small molecule kinase inhibitor selectivity, and the impact that missense mutations have on kinase structure and inhibitor binding. To assess the utility of free energy calculations for predicting selectivity, we performed relative free energy calculations on publicly available congeneric series of ligands on multiple kinase targets. We built a Bayesian graphical model to quantitate the correlation of errors for a given ligand on both target, to interrogate whether any fortuitous

cancellation of errors makes selectivity predictions more accurate than expected. To understand the functional impact of mutations on protein kinase structure and inhibitor binding, we performed massively parallel molecular dynamics simulations of clinically observed hyperactivating mTOR mutations and alchemical free energy calculations on clinical Abl mutations. To rigorously test our predictions, we have developed a panel of kinase expression constructs (now available through AddGene) appropriate for automated, high-throughput expression protocol in *E. coli*. We have demonstrated the utility of these constructs for engineering and expressing clinically observed missense mutations, testing a panel of 96 mutations in Src and Abl kinases gathered from publicly available cancer genomics datasets as well as the MSK-IMPACT clinical sequencing panel, and further expressed a separate panel of 95 clinically relevant Abl mutations. We measured the binding free energies for these clinically relevant Abl mutations for a panel of FDA-approved small molecule kinase inhibitors. Using this dataset as a benchmark, we tested the sensitivity and accuracy of absolute free energy calculations to predict the impact of mutations on inhibitor binding. Taken together, this work provides an assessment of physical modeling for predicting selectivity and resistance in drug design of small molecule kinase inhibitors.

BIOGRAPHICAL SKETCH

Steven was born to Eileen and Joseph Albanese in Wayne, New Jersey in 1991, where he would spend the rest of his childhood. Steven graduated from Wayne Hills High School in 2010, and began college at Cornell University that fall. Steven would graduate *cum laude* with a Bachelor of Arts in Cell Biology and Chemistry in 2014. During his time at Cornell, Steven worked in and completed an honors thesis in the laboratory of Holger Sondermann in the Department of Molecular Medicine, studying the molecular binding partners of Atlastin, a commonly mutated gene in spastic paraplegia. Steven spent two years working as a study group leader for Intro Biology with the Biology Scholars Program. In July 2014, Steven began his first year at Louis V. Gerstner Jr. Graduate School of Biomedical Sciences, achieving a life-long goal of working at Memorial Sloan Kettering Cancer Center.

To my father, whose unwavering support and love inspired me and kept me
going. He is terribly missed.

ACKNOWLEDGEMENTS

Your acknowledgements go here. Make sure it sits inside the brackets.

Table of Contents

Biographical Sketch	v
Dedication	vi
Acknowledgements	vii
Table of Contents	viii
List of Tables	xi
List of Figures	xii
1 Introduction	1
1.1 Perspective	1
1.2 Synopsis	5
2 Predicting the selectivity of small molecule kinase inhibitors	7
2.1 Abstract	7
2.2 Introduction	8
2.2.1 Free energy methods can aid structure-based drug design	8
2.2.2 Selectivity is an important consideration in drug design	9
2.2.3 Use of physical modeling to predict selectivity is relatively unexplored	10
2.2.4 Kinases are an interesting and particularly challenging model system for selectivity predictions	10
2.2.5 Assessing the ability of alchemical free energy methods to predict selectivity	12
2.3 Methods	13
2.3.1 Numerical model of selectivity	13
2.3.2 Structure Preparation	14
2.3.3 Ligand Pose Generation	15
2.3.4 Free Energy Calculations	15
2.3.5 Charge Change Free Energy Calculations	16
2.3.6 Statistical Analysis of FEP+ calculations	16
2.3.7 Quantification of the correlation coefficient ρ	17
2.4 Results	20
2.4.1 Free energy methods can be used to predict the selectivity of a compound	20
2.4.2 Correlation of errors can make selectivity predictions more accurate and speed up ligand optimization	21

2.4.3	The CDK2 and CDK9 experimental dataset demonstrates the difficulty in achieving selectivity for closely related kinases	24
2.4.4	The CDK2 and ERK2 dataset achieves higher levels of selectivity for more distantly related kinases	27
2.4.5	FEP+ calculations show accurate potency predictions for ERK2/CDK2 and larger errors for CDK2/CDK9	30
2.4.6	Free energy calculation errors are correlated, accelerating selectivity optimization	32
2.5	Discussion and Conclusions	35
3	Understanding the functional impact of mTOR clinical kinase mutations using physical modeling	40
3.1	Introduction	40
3.1.1	mTOR forms the catalytic core of protein complexes that control a number of cellular processes	40
3.1.2	mTOR is the targeted by inhibitors with two distinct mechanisms of action	43
3.1.3	mTOR signaling is dysregulated in cancer by hyperactivating missense mutations	47
3.1.4	Using physical modeling to understand the functional impact of mTOR mutations	49
3.2	Methods	50
3.2.1	Molecular dynamics simulations	50
3.2.2	Contact Map Analysis	52
3.2.3	Mean Contact formation over time	52
3.2.4	Alchemical free energy calculations	53
3.3	Results	55
3.3.1	Missense mutations perturb the structure of mTOR kinase domain	55
3.3.2	Missense mutations do not appear to shift the formation of an active kinase domain	58
3.3.3	Missense mutations do not disrupt the formation of inhibitory salt bridges between the kinase and FAT domains	60
3.3.4	Free energy calculations show promise in predicting impact of mutations on small molecule and ATP affinity	62
3.4	Discussion and Conclusions	65
4	Predicting the impact of clinically-observed kinase mutations using physical modeling	67
5	Enabling high-throughput biophysical experiments on clinically-observed mutations	68
5.1	Abstract	68
5.2	Introduction	69
5.3	Results	73

5.3.1	Construct boundary choice impacts Abl kinase domain expression	73
5.3.2	Screen of 96 kinases finds 52 with useful levels of automated <i>E. coli</i> expression	78
5.3.3	High-expressing kinases are folded with a well-formed ATP binding site	84
5.3.4	Fluorescence-based thermostability assay	86
5.3.5	ATP-competitive inhibitor binding fluorescence assay	87
5.3.6	Expressing clinically-derived Src and Abl mutants	90
5.4	Methods	96
5.4.1	Semi-automated selection of kinase construct sequences for <i>E. coli</i> expression	96
5.4.2	Mutagenesis protocol	100
5.4.3	Expression testing	101
5.4.4	Fluorescence-based thermostability assay	102
5.4.5	ATP-competitive inhibitor binding fluorescence assay	104
5.4.6	Large Scale expression and purification protocol for MK14 . .	107
5.5	Discussion	108
5.6	Author Contributions	110
5.7	Acknowledgments	111

List of Tables

5.1	Kinase domain constructs with yields $\geq 2 \mu\text{g}/\text{mL}$ culture for 96-kinase expression screen.	81
5.2	Expression yields for engineered clinical missense mutants of Abl kinase domains with yields $\geq 2 \mu\text{g}/\text{mL}$ culture.	91
5.3	Expression yields for engineered clinical missense mutants of Src kinase domains with yields $\geq 2 \mu\text{g}/\text{mL}$ culture.	92

List of Figures

2.1	Free energy calculations speed up selectivity optimization	23
2.2	A CDK2/CDK9 selectivity dataset from Shao et al., 2013	26
2.3	CDK2 and ERK2 selectivity dataset from Blake et al., 2016	29
2.4	Relative free energy calculations can accurately predict potency, but show larger errors for selectivity predictions.	31
2.5	Correlation in selectivity prediction errors can be used to accelerate selectivity optimization	34
2.6	CDK2 adopts an inactive conformation in the crystal structure used for the CDK2/ERK2 calculations	37
2.7	Correlation coefficient ρ controls the shape of the joint marginal distribution of errors	38
2.8	Correlation reduces the expected error for selectivity predictions	39
3.1	mTOR is an atypical kinase with a number of regulatory domains	42
3.2	mTORC1 integrates signaling from a number of inputs	46
3.3	Hyperactivating mTOR missense mutations have been observed in cancer	48
3.4	Missense mutations can perturb local structure	57
3.5	Missense mutations do not shift the population of the active conformation on a common activation order parameter	59
3.6	Missense mutations in kinase domain do not disrupt interactions between the kinase and FAT domains	61
3.7	Free energy calculations identify potential resistance mutations	64
5.1	Abl kinase domain construct expression screen illustrates high sensitivity to construct boundaries.	76
5.2	Expression yields of Abl kinase domain constructs for all constructs with detectable expression.	77
5.3	Kinome wide search for expressible kinases.	80
5.4	Fluorescence-based thermostability assay demonstrates many high-expressing kinases are well-folded.	85
5.5	Fluorescence emission spectra as a function of the fluorescent ATP-competitive kinase inhibitor bosutinib demonstrates the presence of a well-formed ATP binding pocket.	89
5.6	Expression yields for engineered clinically-derived Src and Abl missense mutants	95

CHAPTER 1

INTRODUCTION

1.1 Perspective

In 2018, cancer accounted for 610,000 deaths and 1.7 million diagnoses in the United States alone [1]. Since the FDA approval of imatinib in 2001, therapeutics targeting kinases now account for over 50% of current cancer drug discovery and close to 30% of total drug discovery efforts [2], with 44 FDA approved small molecule kinase inhibitors (SMKIs) on the market [3]. However, there has been a decrease in productivity using current design strategies, with many drugs failing in late stage clinical trials. By the time a drug fails in Phase III, a typical pharmaceutical company has spent 12 years and almost \$1 billion on development [4]. SMKIs can fail late in the development pipeline for two main reasons: safety issues or lack of efficacy. Tumors have multiple routes to resistance, including target amplification [5, 6], effectively increasing the amount of drug required to get the same level of inhibition. Inhibitor resistance occurs through the presence or upregulation of a redundant pathway [7, 8], mutation of the target kinase [9, 10], activation of downstream kinases [11], or relief of feedback inhibition [12]. On-target toxicity, from inhibition of wild type kinase, can cause efficacy issues by limiting the maximally tolerated dose (MTD). Safety issues arise from adverse events due off-target toxicity, such as gefitinib inhibiting CYP2D68 and causing hepatotoxicity in lung cancer patients, or from the on-target toxicity of inhibiting the wild type kinase [13, 14].

Each kinase inhibitor has a certain selectivity profile, or group of biological targets a molecule binds to and inhibits strongly enough to produce a phenotype. Kinase inhibitors have potential for a great diversity of selectivity—the number of

targets a molecule binds to below a certain K_d threshold. There are more than 520 members of the human kinome [15], each with a highly similar, druggable ATP-binding site [16–20], giving inhibitors targeted to them huge potential for promiscuity, like staurosporine, which inhibits a large percentage of the kinome with high affinity. Even FDA approved drugs have a wide range of selectivities. In a 2011 paper, Davis *et al.*, characterized the interaction of 72 known kinase inhibitors against a panel of 442 kinases [21] using a competitive binding assay. Of the 72 compounds screened, 70% had a K_d less than 3 μM for more than 10% of the 442 distinct kinases screened. While this study confirmed that type II inhibitors, SMKIs that bind an active site adjacent pocket exposed in the ‘DFG-out’ conformation, are more likely to be selective than Type I inhibitors, those that can bind to either the ‘DFG-out’ or ‘DFG-in’ conformation, it also found that there are several type II inhibitors that have low selectivity. Conversely, several Type I inhibitors exhibited a high level of selectivity. This suggests that either binding mode is a viable option when seeking to design a selective inhibitor. Additionally, 17 of the 72 compounds bound to fewer than 5 off-target kinases with affinity comparable to their primary target and also had a K_d less than 3 μM for less than 10% of the assayed kinases. This suggests that it is possible to design compounds that are selective for multiple targets, a strategy that has been suggested as a possible design paradigm termed targeted polypharmacology [11, 22–24].

Current design efforts focus on achieving maximal selectivity for a single target by improving a weak inhibitor through analogue synthesis [25], which is not always rational, or through structure-informed design [25, 26], which is difficult because kinases exist as nodes in complex signaling networks [27, 28], with feedback inhibition and pathway cross-talk complicating the relationship between binding and signaling. This complicates the notion of inhibiting a single kinase to shut-down a pathway, as alleviating negative feedback can lead to re-activation of the

target pathway [27, 29], or lead to activation of a secondary pathway previously regulated by inhibitory cross-talk [27, 30]. Further, tumors can easily evade inhibition of a single target [11] by mutating the target to ablate inhibitor binding; mutating downstream effectors to bypass the inhibited node in the pathway; up-regulating a redundant signaling pathway or branch; or up-regulating the target kinase to increase the amount of drug needed for efficacious inhibition to occur. Rationally designing SMKIs for a given selectivity profile, or to overcome resistance will be essential to continue making advances in the treatment of cancer. Selectivity can be improved by antitargeting kinases closely related to the desired therapeutic target, such as positively designing an inhibitor for EGFR while antitargeting HER2. Multitarget design could also reduce on-target toxicity, thereby improving the therapeutic window, by targeting the oncogenic mutated kinase and antitargeting the wild type kinase. This could improve upon the success of certain EGFR inhibitors such as gefitinib and erlotinib [31–33] or aid the development of second generation inhibitors for use in treating patients with clinically-acquired resistance mutations [10, 34, 35], such as the ALK inhibitor alectinib [36]

A further factor complicating the design of small molecule kinase inhibitors is the proliferation of mutations observed in the clinic. Next-generation sequencing has enabled generation of massive datasets rich with missense alterations in kinases observed directly in the clinic [37–39], and has been particularly transformative in the field of oncology. While this technology has drastically advanced our understanding of disease, it presents a problem for rational drug design. While some missense mutations are highly recurrent and have been characterized clinically or biochemically, a long tail of rare mutations accounts for the majority of clinically observed mutations, leaving clinicians, researchers, and chemists without an understanding of whether these mutations might be activating and drive cancer, or cause resistance to a previously-developed compound. Understanding

the impact of these mutations will enable the development of next-generation inhibitors that overcome resistance and have lower-levels of on-target toxicity by sparing the wild-type form of the kinase in favor of the oncogenic mutant kinase.

Physical modeling can be used to characterize the impact of these rare mutations, as well as enable the design of inhibitor selectivity. Molecular dynamics have been successfully applied to a number of drug discovery projects [40], such as for identifying allosteric binding sites [41] or enabling more accurate virtual docking screens [42] through the relaxed complex scheme (RCS) [41, 43]. While these methods have lead to successful development of inhibitors for HIV integrase [44] and FKBP [41] as well as some kinase [45], docking scoring algorithms focus on enrichment [46] and do not correlate well with ligand binding affinity [47]. Alchemical free energy calculations allow for prediction of ligand binding free energies, including all enthalpic and entropic contributions [48–50]. Advances in atomistic molecular mechanics forcefields and free energy methodologies [51–54] have allowed free energy methods to reach high levels of accuracy for predicting ligand potencies [55]. Free energy methods have been applied prospectively to develop inhibitors for Tyk2 [56], Syk [57], BACE1 [58], GPCRs [59], and HIV protease [60]. While predicting affinity with free energy calculations has been well-studied, the utility of these methods for predicting selectivity is understudied. Early work on using free energy calculations to understand the selectivity of imatinib for Abl over Src [61, 62] and predicting the selectivity of three inhibitors within the bromodomain family [63] is promising that these methods can be used to design the selectivity of inhibitors.

Physical models can also be used to understand the impact of clinical mutations on the structure and function of proteins, enabling the design and selection of inhibitors for the era of personalized medicine. Molecular Dynamics simula-

tions have been applied to understanding the mechanism of mutations on the activity [64, 65] and ligand binding [66] of EGFR. They have also been applied to understanding missense mutations in p53 [67], CLIC2 [68], and opsin [69]. Alchemical free energy calculations have been applied to several widespread studies of protein thermostability [70–72] as well as understanding how gatekeeper mutations impact inhibitor binding [73].

As will be discussed at length in this work, physical modeling can enable the predicting of selectivity and understanding the impact mutations have on protein structure, function, and inhibitor binding. These methods can greatly aid the design of small molecule kinase inhibitors that will overcome the complex biology of kinases and overcome resistance, both of which limit the benefit patients treated with current compounds receive.

1.2 Synopsis

This thesis is organized as follows. Chapter 2 contains a manuscript presenting a study of the utility of free energy methods to predict selectivity, using small molecule kinase inhibitors as a clinically-relevant and particularly challenging test system. It will present a numerical model for the speedup that can be expected when designing the selectivity of an inhibitor using free energy methods. Chapter 3 contains work done to understand the impact of clinical mTOR mutations on protein structure and understand the role these mutations play in activating the kinase, as well as the limitations of this type of analysis. Chapter 4 partially presents work from a manuscript published in *Communications Biology* that uses alchemical free energy methods to calculate the impact of clinical mutations on inhibitor binding affinities for Abl kinase. This work highlights the benefit of using physical

models to compute physically-meaningful and testable quantities, as well as highlights some of the challenges and limitations in using publically available data to benchmark free energy calculations. Finally, Chapter 5 contains a manuscript published in *Biochemistry* that presents work done to enable the generation of high-quality binding affinity data, that is critical for testing the accuracy and utility of free energy methods in drug discovery.

CHAPTER 2

PREDICTING THE SELECTIVITY OF SMALL MOLECULE KINASE INHIBITORS

The work in this chapter is in preparation, and will be submitted as follows.

Is structure based drug design ready for selectivity optimization?

Steven K. Albanese^{1,2}, John D. Chodera², Simon Peng³, Robert Abel³, Lingle Wang^{3,*}

¹ *Louis V. Gerstner, Jr. Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, NY 10065*

² *Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065*

³ *Schrödinger, New York, NY 10036*

* Corresponding Author

2.1 Abstract

Alchemical free energy calculations are now widely used to drive or maintain potency in small molecule lead optimization, where the binding affinity to a protein target can be computed—in well-behaved cases—to roughly 1 kcal/mol inaccuracy, which is believed to primarily stem from force field errors. Despite this, the potential to use free energy calculations to drive optimization of compound *selectivity* among two similar targets has been relatively unexplored. In the most optimistic scenario, the similarity of binding sites might lead to a fortuitous cancellation of force field errors and allow selectivity to be predicted more accurately than affinity. Here, we assess the accuracy with which selectivity can be predicted in the context of small molecule kinase inhibitors, considering the very similar binding sites of human kinases CDK2 and CDK9, as well as another series of ligands attempting to achieve selectivity between the more distantly related kinases CDK2 and ERK2. Using a novel Bayesian analysis

approach, we separate force field error from statistical error and quantify the correlation in force field errors between selectivity targets. We find that, in the closely related CDK2/CDK9 case, a high correlation in force field errors suggests free energy calculations can have significant impact in aiding chemists in achieving selectivity, while in more distantly related kinases (CDK2/ERK2), limited correlation in force field errors reduces the ability for free energy calculations to aid selectivity optimization. In both cases, the correlation in force field error suggests that longer simulations are beneficial to properly balance statistical error with systematic error to take full advantage of the increase in accuracy in selectivity prediction possible due to fortuitous cancellation of error.

2.2 Introduction

2.2.1 Free energy methods can aid structure-based drug design

Free energy methods have proven useful in aiding structure-based drug design by driving the optimization or maintenance of potency in lead optimization. Alchemical free energy calculations allow for prediction of ligand binding free energies, including all enthalpic and entropic contributions [48]. Advances in atomistic molecular mechanics forcefields and free energy methodologies [51–54] have allowed free energy methods to reach a level of accuracy sufficient for predicting ligand potencies [55]. Free energy methods have been applied prospectively to develop inhibitors for Tyk2 [56], Syk [57], BACE1 [58], GPCRs [59], and HIV protease [60]. A recent large-scale review found that the use of FEP+ [74] to predicting potency for 92 different projects and 3021 compounds found a median RMSE of 1 kcal/mol [75].

2.2.2 Selectivity is an important consideration in drug design

In addition to maintaining or optimizing potency, free energy methods can be applied to predicting the selectivity of a ligand between two or more targets. Selectivity is an important property to consider in drug development, either in the pursuit of a maximally selective inhibitor [25, 26] or in pursuit a polypharmacological agent [11, 22–24, 76], to avoid on-target toxicity (arising from inhibition of the intended target) [13] and off-target toxicity (arising from inhibition of unintended targets) [14, 77]. In either paradigm, considering the selectivity of a compound is complicated by the biology of the target. For example, kinases exist as nodes in complex signaling networks [27, 28] with feedback inhibition and cross-talk between pathways. Careful consideration of which off-targets are being inhibited can avoid off-target toxicity due to alleviating feedback inhibition and inadvertently reactivating the targeted pathway [27, 28], or the upregulation of a secondary pathway by alleviation of cross-talk inhibition [12, 30]. Off-target toxicity can also be caused by inhibiting unrelated targets, such as gefitinib, an EGFR inhibitor, inhibiting CYP2D6 [77] and causing hepatotoxicity in lung cancer patients. In a cancer setting, on-target toxicity can be avoided by considering the selectivity for the oncogenic mutant form of the kinase over the wild type form of the kinase [78–80], demonstrated by number of first generation EGFR inhibitors. Selectivity considerations can also lead to beneficial effects: Imatinib, initially developed to target BCR-Abl fusion proteins, is also approved for treating gastrointestinal stromal tumors (GIST) [81] due to its activity against receptor tyrosine kinase KIT.

2.2.3 Use of physical modeling to predict selectivity is relatively unexplored

While predicting selectivity is important for drug discovery, but the utility of free energy methods for predicting this property is relatively unexplored. If there is fortuitous cancellation of errors for closely related systems, free energy methods may be much more accurate than expected given the errors made in predicting the potency for each individual target. The selectivity of Imatinib for Abl kinase over Src [61, 62] and within a family of non-receptor tyrosine kinases [82] has been studied extensively using molecular dynamics and free energy calculations. This work focuses on understanding the role reorganization energy plays in the exquisite selectivity of imatinib for Abl over Src despite high similarity between cocrystallized binding mode and kinase conformations, and does not touch on the evaluation of the accuracy of these methods, or their application to drug discovery on congeneric series of ligands. Previous work predicting the selectivity of three bromodomain inhibitors across the bromodomain family achieved promising accuracy for single target potencies of roughly 1 kcal/mol, but does not explicitly evaluate any selectivity metrics [63] or look at correlation in the errors made for each bromodomain.

2.2.4 Kinases are an interesting and particularly challenging model system for selectivity predictions

Kinases are a useful model system to work with for assessing the utility of free energy calculations to predict selectivity. With the approval of imatinib for the treatment of chronic myelogenous leukemia in 2001, targeted small molecule

kinase inhibitors (SMKIs) have become a major class of therapeutics in treating cancer and other diseases. Currently, there are 44 FDA-approved SMKIs [3], and it is estimated that kinase targeted therapies account for as much as 50% of current drug development [83], with many more compounds currently in clinical trials. While there have been a number of successes, the current stable of FDA-approved kinase inhibitors targets only a small number of kinases implicated in disease, and the design of new selective kinase inhibitors remains a significant challenge. Achieving desired selectivity profiles is particularly difficult for kinase targets, making them a system where physical modelling has the potential for a large impact. Achieving selective inhibition of kinases is challenging as there are more than 518 protein kinases [15, 84] with a highly conserved ATP binding site that is targeted by the majority of SMKIs [85]. While kinase inhibitors have been designed to target kinase-specific subpockets and binding modes to achieve selectivity [16–19, 86, 87], previous work has shown that both Type I (binding to the active, DFG-in conformation) and Type II (binding to the inactive, DFG-out conformation) inhibitors display a wide variety of selectivities [21, 88], often exhibiting significant binding a number of other targets in addition to their primary target. Even FDA-approved inhibitors—often the result of extensive drug development programs—bind to a large number of off-target kinases [89]. Kinases are also targets of interest for developing polypharmacological compounds, or inhibitors that are specifically designed to inhibit multiple kinase targets. Resistance to MEK inhibitors in KRAS-mutant lung and colon cancer has been shown to be driven by HER3 upregulation [90], providing rational for dual MEK/ERBB family inhibitors. Similarly, combined MEK and VEGFR1 inhibition has been proposed as a combinatorial approach to treat KRAS-mutant lung cancer [91]. Developing inhibitors with the desired polypharmacology means navigating more complex selectivity profiles. In

well-behaved kinase systems, free energy calculations potency predictions have achieved mean unsigned errors of less than 1.0 kcal/mol [56, 74], suggesting that kinases can be computationally tractable as well as clinically interesting.

2.2.5 Assessing the ability of alchemical free energy methods to predict selectivity

We anticipate difficulty in predicting selectivity if the errors in the alchemical free energy calculations for two targets are largely uncorrelated, or even anticorrelated. However, correlation in the forcefield errors of the free energies for the two targets could lead to a fortuitous cancellation of errors in predicting the selectivity between targets, making selectivity predictions *more* accurate than potency predictions. Such correlation could occur because the same chemical elements appear in the ligand and in highly related binding sites. Here, we investigate the magnitude of this correlation (ρ) and the utility of alchemical free energy calculations for the prediction of selectivity, hereafter taken to mean the $\Delta\Delta G$ in binding free energies of the same compound for two targets. We employed state of the art relative free energy calculations [74, 75] to predict the selectivities of two different congeneric ligand series [92, 93], as well as present a simple numerical model to quantify the potential speed up in selectivity optimization expected for different combinations of per target errors and correlation coefficient values. To tease out the effects of a limited number of experimental measurements, we develop a new Bayesian approach to quantify the uncertainty in the correlation coefficient in the predicted change in selectivity on ligand modification, incorporating all sources of uncertainty and correlation in the computation to separate statistical from force field error. We find that in the

closely related systems of CDK2 and CDK9, a high correlation of force field errors suggests that free energy methods can have a significant impact on speeding up selectivity optimization. In the more distantly related case (CDK2/ERK2), limited correlation hampers the ability for free energy methods to speed up selectivity optimization.

2.3 Methods

2.3.1 Numerical model of selectivity

To model the impact correlation would have on the expected uncertainty for selectivity predictions, $\sigma_{selectivity}$ was calculated using Equation 2.1 for 1000 evenly spaced values of the correlation coefficient (ρ) from 0 to 1, for a number of combinations of per target errors ($\sigma_{target1}$ and $\sigma_{target2}$)

$$\sigma_{selectivity} = \sqrt{\sigma_{target1}^2 + \sigma_{target2}^2 - 2\rho\sigma_{target1}\sigma_{target2}} \quad (2.1)$$

The speed up in selectivity optimization that could be expected from using free energy calculations of a particular per target error ($\sigma_{selectivity}$) was quantify as follows using NumPy (v 1.14.2). An original, true distribution for the changed in selectivity of 200000000 new compounds proposed with respected to a reference compound was modeled as a normal distribution centered around 0 with a standard deviation of 1 kcal/mol. This assumption was made on the basis that the majority of selectivity is driven by the scaffold, and R group modifications will do little to drive changes in selectivity. The 1 kcal/mol distribution is supported by the standard deviations of the selectivity in the experimental

datasets referenced in this work, which are all less than, but close, to 1 kcal/mol.

Each of these proposed compounds were "screened" by a free energy calculation technique with a per target error (σ_{target}) of 1 kcal/mol [74] and a specified correlation coefficient ρ . A $\sigma_{selectivity}$ was calculated according to Equation 2.1. The noise of the computational method was modeled as a normal distribution centered around 0 with a standard deviation of $\sigma_{selectivity}$ and added to the "true" change in selectivity. Any compound predicted to have an improvement in selectivity of 1.4 kcal/mol (1 log unit) would then be made and have its selectivity experimental measured. The speedup value for each value of ρ is calculated as the proportion of compounds made with a true selectivity gain of 1.4 kcal/mol divided by the proportion of compounds with a 1.4 kcal/mol improvement in the original distribution, where all of the compounds were made.

Finally, this process was repeated for a 100x (2.8 kcal/mol, 2 log unit) selectivity optimization and 50 linearly spaced values of the correlation coefficient (ρ) between 0 and 1, for four values of $\sigma_{selectivity}$ and 40000000 compounds in the original distribution.

2.3.2 Structure Preparation

Structures from the Shao [92] and Hole [94], and Blake [93] papers were downloaded from the PDB [95], selecting structures with the same co-ligand crystallized. For the Shao dataset, 4BCK (CDK2) and 4BCI (CDK9) were selected, which have ligand 12c cocrystallized. For the Blake dataset, 5K4J (CDK2) and 5K4I (ERK2) were selected, cocrystallized with ligand 21. The structures were prepared using Schrodinger's Protein Preparation Wizard [96] (release 2017-3). This pipeline modeled in internal loops and missing atoms, added hydrogens at

the reported experimental pH (7.0 for the Shao dataset, 7.3 for the Blake dataset) for both the protein and the ligand. All crystal waters were retained. The ligand was assigned protonation and tautomer states using Epik at the experimental pH \pm 2, and hydrogen bonding was optimized using PROPKA at the experimental pH \pm 2. Finally, the entire structure was minimized using OPLS3 with an RMSD cutoff of 0.3.

2.3.3 Ligand Pose Generation

Ligands were extracted from the publication entries in the BindingDB as 2D SMILES strings. 3D conformations were generated using LigPrep with OPLS3 [97]. Ionization state was assigned using Epik at experimental pH \pm 2. Stereoisomers were computed by retaining any specified chiralities and varying the rest. The tautomer and ionization state with the lowest epik state penalty was selected for use in the calculation. Ligand poses were generated by first aligning to the co-crystal ligand using the Largest Common Bemis-Murcko scaffold with fuzzy matching (Schrodinger 2017-4). Ligands that were poorly aligned or failed to align were then aligned using Maximum Common Substructure (MCSS). Finally, large R-groups were allowed to sample different conformations using MM-GBSA with a common core restrained. VSGB solvation model was used with the OPLS3 forcefield. No flexible residues were defined for the ligand.

2.3.4 Free Energy Calculations

The FEP+ panel (Maestro release 2017-4) was used to generate perturbation maps. Neutral perturbations were run for 15ns per replica, using an NPT ensemble and

water buffer size of 5. A GCMC solvation protocol was used to sample buried water molecules in the binding pocket prior to the calculation, which discards any retained crystal waters.

2.3.5 Charge Change Free Energy Calculations

For ligands where a protonation state change was expected to be relevant to binding based on a small state penalty, Jaguar pKa prediction calculations [98] were run to identify protonation state changes with pKas within 1 log unit of the experimental pH. The predicted pKas for one ligand (Shao 12b, 7.84) was within this range. To account for this, a pKa correction was performed. For this ligand, a separate perturbation map containing ligands 12a, 12c, 12b (neutral) and 12b (charged) was run for 30ns per replica using a post-calculation Coulombic charge correction. Each charge change perturbation edge was run with a 150mM NaCl concentration. The pKa correction was performed using Equation 2.2:

$$\Delta\Delta G_{corrected} = \Delta\Delta G_{uncorrected} - RT \log \left(\frac{10^{pK_a-pH} + 1}{e^{\frac{\Delta G_{neutral}-\Delta G_{charged}}{RT}} * (10^{pK_a-pH} + 1)} \right) \quad (2.2)$$

$\Delta\Delta G$ for each edge in perturbation map with 12a, 12c and 12b (neutral) was updated using the correction above and merged into the final map.

2.3.6 Statistical Analysis of FEP+ calculations

Each FEP+ calculation has a reported mean unsigned error (MUE) and root mean squared error (RMSE) with a bootstrapped 95% confidence interval. The MUE

was calculated according Equation 2.3, while the RMSE was calculated according to Equation 2.4.

$$\text{MUE} = \frac{\sum_0^n |\Delta G_{\text{calc}} - \Delta G_{\text{exp}}|}{n} \quad (2.3)$$

$$\text{RMSE} = \frac{\sqrt{\sum_0^n (\Delta G_{\text{calc}}^2 - \Delta G_{\text{exp}}^2)}}{n} \quad (2.4)$$

Each RMSE and MUE is reported with a 95% confidence interval calculated from 10000 replicates of a choose-one-replace bootstrap protocol on the ΔG values reported to account for the finite sample size of the ligands. The code used to bootstrap these values is available on github:

<https://github.com/choderalab/selectivity>

2.3.7 Quantification of the correlation coefficient ρ

To quantify ρ , we built a Bayesian graphical model using pymc3 (v. 3.5) [99] and theano (v 1.0.3) [100], which has been made available on Github. For each phase (complex and solvent), the absolute free energy (G) of ligand i was treated as a normal distribution (Equation 2.5). For each set of calculations, one ligand was chosen as the reference, and pinned to 0, with a standard deviation of 1 kcal/mol in order to improve the efficiency of sampling from the model.

$$G_{i,\text{target}}^{\text{phase}} = \mathcal{N}(\mu = 0, \text{sd} = 25.0 \text{ kcal/mol}) \quad (2.5)$$

For each edge of the FEP map (ligand $i - j$ ligand j), there is a contribution from

dummy atoms, that was modeled as in Equation 2.6.

$$c_{i,j} = \mathcal{N}(\mu = 0, \text{ } sd = 25.0 \text{ kcal/mol}) \quad (2.6)$$

The model was restrained by including data from the FEP+ calculation.

$$\Delta G_{\text{phase}, ij, \text{target}}^{\text{BAR}} = \mathcal{N}(G_{j,\text{target}}^{\text{phase}} - G_{i,\text{target}}^{\text{phase}}, \delta^2 \Delta G_{\text{phase}, ij, \text{target}}^{\text{BAR}}, \text{observed} = \Delta G_{\text{phase}, ij, \text{target}}^{\text{calc}}) \quad (2.7)$$

Where $\delta^2 \Delta G_{\text{phase}, ij, \text{target}}^{\text{BAR}}$ is the reported BAR uncertainty from the calculation, and $\Delta G_{\text{phase}, ij, \text{target}}^{\text{calc}}$ is the BAR estimate of the free energy for the perturbation between ligands i and j in a given phase.

From this, we can calculate the $\Delta\Delta G^{\text{FEP}}$ for each edge as in Equation 2.8:

$$\Delta\Delta G_{\text{target}, ij}^{\text{FEP}} = \Delta G_{\text{complex}, ij, \text{target}}^{\text{BAR}} - \Delta G_{\text{solvent}, ij, \text{target}}^{\text{BAR}} \quad (2.8)$$

To model the way an offset is calculated for the ΔG reported by the FEP+ panel in Maestro:

$$\text{offset} = \frac{\sum^n G_{i,\text{target}}^{\text{complex}} - G_{i,\text{target}}^{\text{solvent}}}{n} - \frac{\sum^n \Delta G_i^{\text{exp}}}{n} \quad (2.9)$$

The offset was added to each ΔG_i^{BAR} to calculate ΔG_i^{sch} .

The experimental binding affinity was treated as a true value ($\Delta G_{i,\text{target}}^{\text{true}}$) corrupted by experimental uncertainty, which is assumed to be 0.3 kcal/mol [55], with the values reported in the papers ($\Delta G_{i,\text{target}}^{\text{obs}}$) treated as observations from this distribution (Equation 2.10)

$$\Delta G_{i,target}^{exp} = \mathcal{N}(\Delta G_{i,target}^{true}, 0.3 \text{ kcal/mol}, \text{ observed} = \Delta G_{i,target}^{obs}) \quad (2.10)$$

$\Delta G_{i,target}^{true}$ was assigned a weak normal prior, as in equation 2.11.

$$\Delta G_{i,target}^{true} = \mathcal{N}(0, 50 \text{ kcal/mol}) \quad (2.11)$$

The error for a given ligand was calculated as in Equation 2.12.

$$\epsilon_i = \Delta G_i^{sch} - \Delta G_i^{true} \quad (2.12)$$

From these ϵ values, we calculated the correlation coefficient, ρ as in Equation 2.13.

$$\rho = \frac{cov(\epsilon_{target1}, \epsilon_{target2})}{\sigma_{target1}\sigma_{target2}} \quad (2.13)$$

Where σ is the standard deviation of ϵ . To quantify ρ for the CDK2/ERK2 calculations, the default NUTS sampler with jitter+adapt_diag initialization, 1000 tuning steps, and a target accept probability of 0.8 was used to draw 10000 samples from the model. The CDK2/CDK9 model was sampled 20000 times using default NUTS sampler with jitter+adapt_diag initialization and 3000 tuning steps.

2.4 Results

2.4.1 Free energy methods can be used to predict the selectivity of a compound

While ligand potency for a single target is often quantified as a free energy of binding ($\Delta G_{binding}$), there are a number of different metrics for quantifying the selectivity of a compound [101, 102]. Here, we propose a more granular view of selectivity: the change in free energy of binding for a given ligand between two different targets ($\Delta\Delta G_{selectivity}$), which can be calculated as in Equation 2.14. $\Delta\Delta G_{selectivity}$ is a useful measure of compound selectivity once a single, or small panel, of off-targets have been identified.

$$\Delta\Delta G_{selectivity} = \Delta G_{binding, \text{target } 2} - \Delta G_{binding, \text{target } 1} \quad (2.14)$$

To predict the $\Delta\Delta G_{selectivity}$ of a compound, we developed a protocol that uses a relative free energy calculation (FEP+) [74] to run a map of perturbations between ligands in a congeneric series, as described in depth in the methods section. The calculation is repeated for each target of interest, with identical perturbations (edges) between each ligand (nodes). Each edge represents a relative free energy calculation that quantifies the $\Delta\Delta G$ between the ligands, or nodes. By using provided experimental data, we can convert the $\Delta\Delta G$ from each edge to a single potency value for each value against that target (ΔG_{target}). From this sets of calculations, we can calculate a $\Delta\Delta G_{selectivity}$ for each ligand given two targets of interest. Previous work shows that FEP+ can achieve an accuracy (σ_{target}) of roughly 1 kcal/mol when predicting potency, which is a combination of

systematic forcefield and random stastistical error [74]. However, it is possible that the forcefield component of that error may fortuitously cancel when computing $\Delta\Delta G_{selectivity}$, leading to a selectivity uncertainty ($\sigma_{selectivity}$) that is lower than would be expected.

2.4.2 Correlation of errors can make selectivity predictions more accurate and speed up ligand optimization

To demonstrate the potential impact correlation has on the uncertainty of selectivity predictions ($\sigma_{selectivity}$) using alchemical free energy techniques, we created a simple numerical model following equation 2.1, which takes into account each of the per target errors expected from the methodology as well as the correlation in those errors. As seen in Figure 2.1A, if the per target erorrs (σ_1 and σ_2) are the same, $\sigma_{selectivity}$ approaches 0 as the correlation coefficient (ρ) approaches 1. If the error for the free energy method is not the same, $\sigma_{selectivity}$ gets smaller but approachs a non-zero value as ρ approaches 1. To quantify the expected speedup in selectivity optimization, we modeled the change in selectivity with respect to a reference compound for a number of compounds a medicinal chemist might suggest as a normal distribution centered around 0 with a standard deviation of 1 kcal/mol (Figure 2.1B, black curve), reflecting that most proposed changes would not drive large changes in selectivity. Then, suppose that each compound is screened computationally with a method free energy methodology with a per target (σ_{target}) error of 1 kcal/mol, and all compounds predicted to have a 1.4 kcal/mol improvement in selectivity are synthesized and experimentally tested (Figure 2.1B, colored curves). The fold-change in the proportion of compounds that are made that have a true 1.4 kcal/mol

improvement in selectivity compared to the original distribution can be calculated as a surrogate for the expected speedup. For a 1.4 kcal/mol selectivity improvement threshold (1 log unit), a correlation of 0.5 gives an expected speed up of 4.1x, which can be interpreted as 4.1x fewer compounds needing to be made before achieving a 1 log unit improvement in selectivity. This process can be extended for the even more difficult proposition of achieving a 2 log unit improvement in selectivity (Figure 2.1C), where 200-300x speedups can be expected, depending on σ_{target} for the free energy methodology.

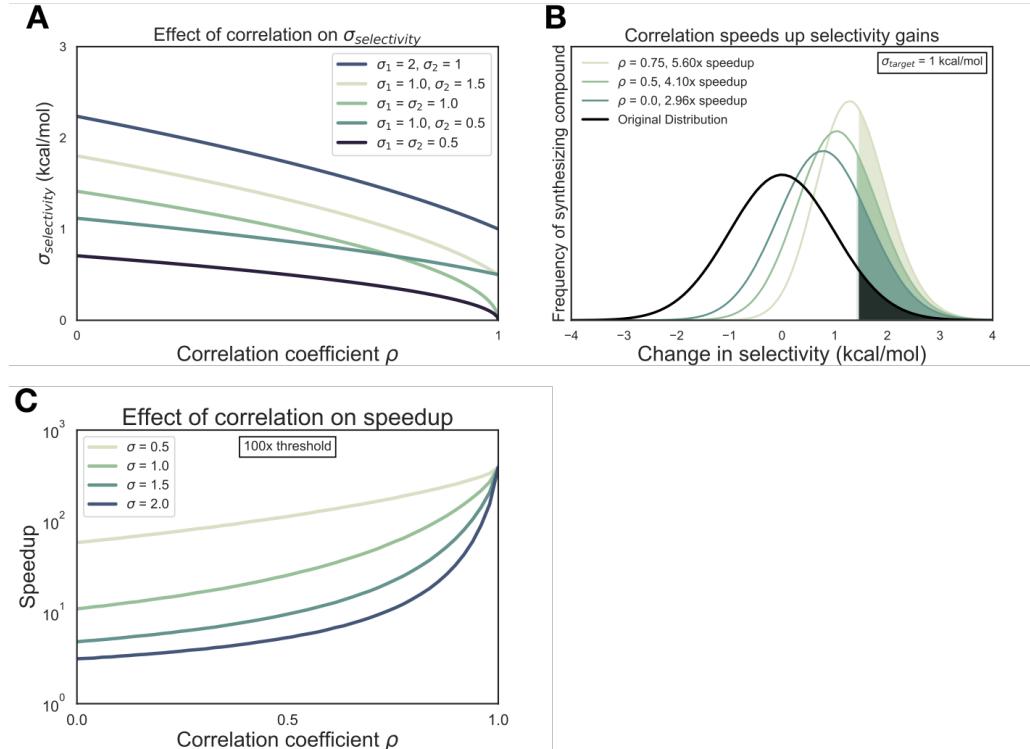


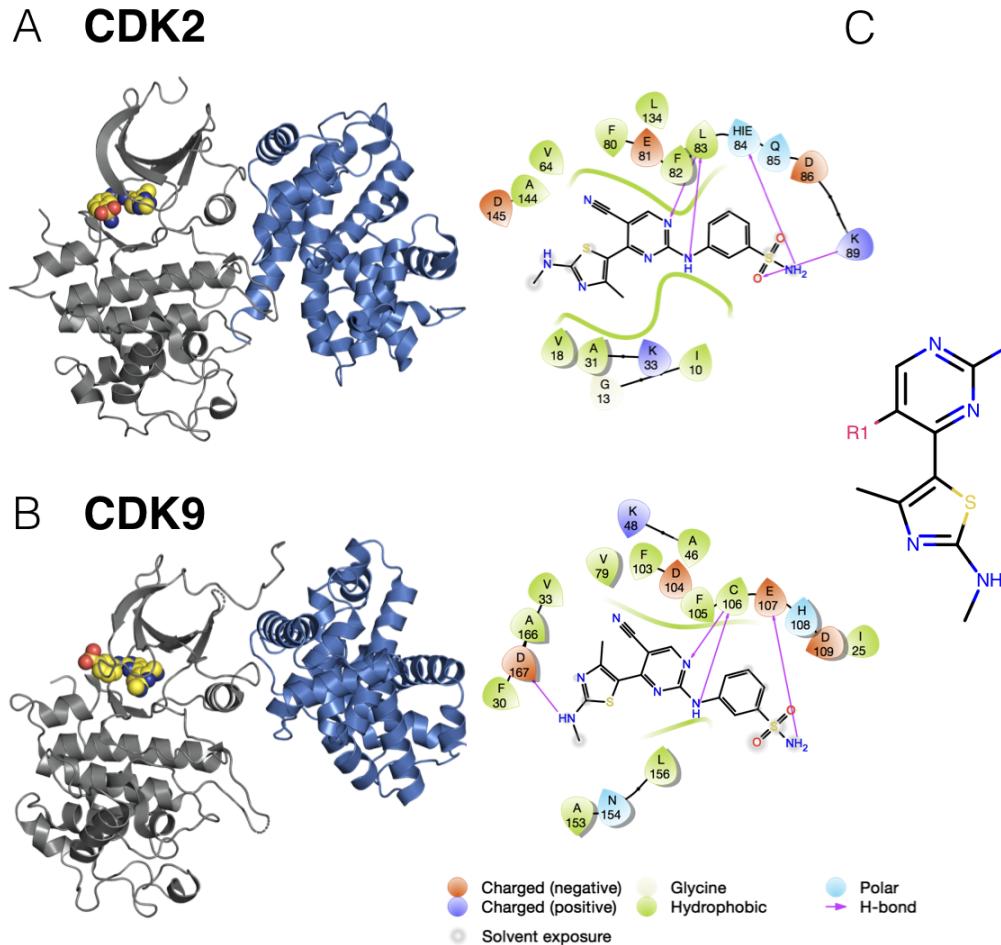
Figure 2.1: **Free energy calculations speed up selectivity optimization** (A) The effect of correlation on expected errors for predicting selectivity ($\sigma_{selectivity}$) in kcal/mol. Each curve represents a different combination of target errors (σ_1 and σ_2). (B) The change in selectivity for molecules proposed by medicinal chemists optimizing a lead candidate can be modeled by a normal distribution centered on 0 with a standard deviation of 1 kcal/mol (black curve). Each green curve corresponds to the distribution of compounds made after screening for a 1 log unit (1.4 kcal/mol) improvement in selectivity with a free energy methodology with a 1 kcal/mol per target error and a particular correlation. The shade region of each curve corresponds to the compounds with a real 1 log unit improvement in selectivity. The speed up is calculated as the ratio of the percentage of compounds made with a real 1 log unit improvement to the percentage of compounds that would be expected in the original distribution. (C) The speedup (y-axis, log scale) expected for 100x (2 log units, 2.8 kcal/mol) selectivity optimization as a function of correlation coefficient ρ . Each curve corresponds to a different σ_{target} value.

2.4.3 The CDK2 and CDK9 experimental dataset demonstrates the difficulty in achieving selectivity for closely related kinases

To begin quantifying the correlation of errors in free energy predictions for selectivity, we set out to gather datasets that met a number of criteria. We looked for datasets that contained binding affinity data for a number of kinase targets and ligands, as well as having crystal structures for each target with the same co-crystallized ligand. For the CDK2/CDK9 dataset [92], ligand 12c was cocrystallized with CDK2/cyclin A (Figure 2.2A, left) and CDK9/cyclin T (Figure 2.2B, left), work that was published in a companion paper [94]. In both CDK2 and CDK9, ligand 12c forms relatively few hydrogen bond interactions with the kinase. Each kinase forms a set of hydrogen bonds between the ligand scaffold and a hinge residue (C106 in CDK9 and L83 in CDK2) that is conserved across all of the ligands in this series. CDK9, which has slightly lower affinity for ligand 12c (Figure 2.2C, right), forms a lone interaction between the sulfonamide of ligand 12c and residue E107. On the other hand, CDK2 forms interactions between the sulfonamide of ligand 12c and residues K89 and H84. The congeneric series of ligands contains a number of challenging perturbations, particularly at substituent point R3 (Figure 2.2C, left). Ligand 12i also presented a challenging perturbation, moving the 1-(piperazine-1-yl)ethanone from the *meta* to *para* location.

This congeneric series of ligands also highlights two of the challenges of working from publicly available data. First, the dynamic range of selectivity is incredibly narrow, with a mean $\Delta\Delta G_{selectivity}$ (CDK9 - CDK2) of only -0.65 kcal/mol, and a standard deviation of 0.88 kcal/mol. Additionally, experimental uncertainties are

not reported for the experimental measurements. Thus, for this and subsequent sets of ligands, the experimental uncertainty is assumed to be 0.3 kcal/mol based on previous work done to summarize uncertainty in experimental data [55, 103].

**C**

Ligand	R1	R2	R3	ΔG CDK2 (kcal/mol)	ΔG CDK9 (kcal/mol)	$\Delta \Delta G$ (kcal/mol)
12a	CN	H		-12.27	-11.21	1.06
12b	OH	H		-7.23	-8.22	-0.99
12c	CN	H		-11.45	-11.21	0.24
12e	F	H		-11.62	-11.45	0.17
12f	Cl	H		-10.91	-10.85	0.06
12g	Methyl	H		-10.18	-11.32	-1.14
12h	Ethyl	H		-8.28	-9.56	-1.28
12j	CN	H		-10.04	-11.12	-1.08
12l	CN		H	-10.34	-10.44	-0.1
12n	CN	H		-10.06	-10.97	-0.91
12o	F	H		-10.06	-11.12	-1.06
12q	F	H		-10.91	-11.62	-0.71
12t	CN	H		-9.38	-11.12	-1.74
1a	H	H		-11.62	-11.86	-0.24
1b	H	H		-11.45	-11.86	-0.41

Figure 2.2: A CDK2/CDK9 selectivity dataset from Shao *et al.*, 2013 (A) (left) Crystal Structure (4BCK)[94] of CDK2 (gray ribbon) bound to ligand 12c (yellow spheres). Cyclin A is shown in blue ribbon (right) 2D ligand interaction map of ligand 12c in the CDK2 binding site. (B) (left) Crystal structure of CDK9 (4BCI)[94] (gray ribbon) bound to ligand 12c (yellow spheres). Cyclin T is shown in blue ribbon. (right) 2D ligand interaction map of ligand 12c in the CDK9 binding site. (C) (left) 2D structure of the common scaffold for all ligands in congeneric ligand series 12 from the publication (right) A table summarizing all R group substitutions as well as the published experimental binding affinities and selectivities[92].

2.4.4 The CDK2 and ERK2 dataset achieves higher levels of selectivity for more distantly related kinases

The CDK2/ERK2 dataset from Blake *et al.*, 2016 also met the criteria described above. Crystal structures for both CDK2 (Figure 2.3A, top) and ERK2 (Figure 2.3B, top) were available with ligand 22 co-crystallized. Of note, CDK2 was not crystallized with cyclin A, despite cyclin A being included in the affinity assay reported in the paper [93]. CDK2 adopts a DFG-in conformation with the α C helix rotated out, away from the ATP binding site and breaking the conserved salt bridge between K33 and E51 (Supp. Figure 2.6A), indicative of an inactive kinase [18, 104]. By comparison, the CDK2 structure from the CDK2/CDK9 dataset adopts a DFG-in conformation with the α C helix rotated in, forming the ionic bond between K33 and E51 indicative of an active kinase, due to allosteric activation by cyclin A. While missing cyclins have caused problems for free energy calculations in prior work, it is possible that the fully active conformation contributes equally to binding affinity for all of the ligands in the series, and the high accuracy of the potency predictions (Figure 2.4, top left) is the result of fortuitous cancellation of errors. The binding mode for this series is similar between both kinases. There is a set of conserved hydrogens bonds between the scaffold of the ligand and the backbone of one of the hinge residues (L83 for CDK2 and M108 for ERK2). The conserved lysine (K33 for CDK2 and K54 for ERK2), normally involved in the formation of a ionic bond with the α C helix, forms a hydrogen bond with the scaffold (Figure 2.4A and 2.4B, bottom) in both CDK2 and ERK2. However, in the ERK2 structure, the hydroxyl engages a crystallographic water as well as N154 in a hydrogen bond network that is not present in the CDK2 structure. The congenic ligand series features a single substituent point, with the R groups exposed to the solvent. This helps explain the

I probably need to either remove the reference to how closely related the kinases are, or come up with a way to quantify that - SKA

is there a good citation for this?

extremely narrow distribution of selectivities, with a mean selectivity of -1.74 kcal/mol (ERK2 - CDK2) and standard deviation of 0.56 kcal/mol. This suggests that the selectivity is largely driven by the scaffold and unaffected by the R group substitutions.

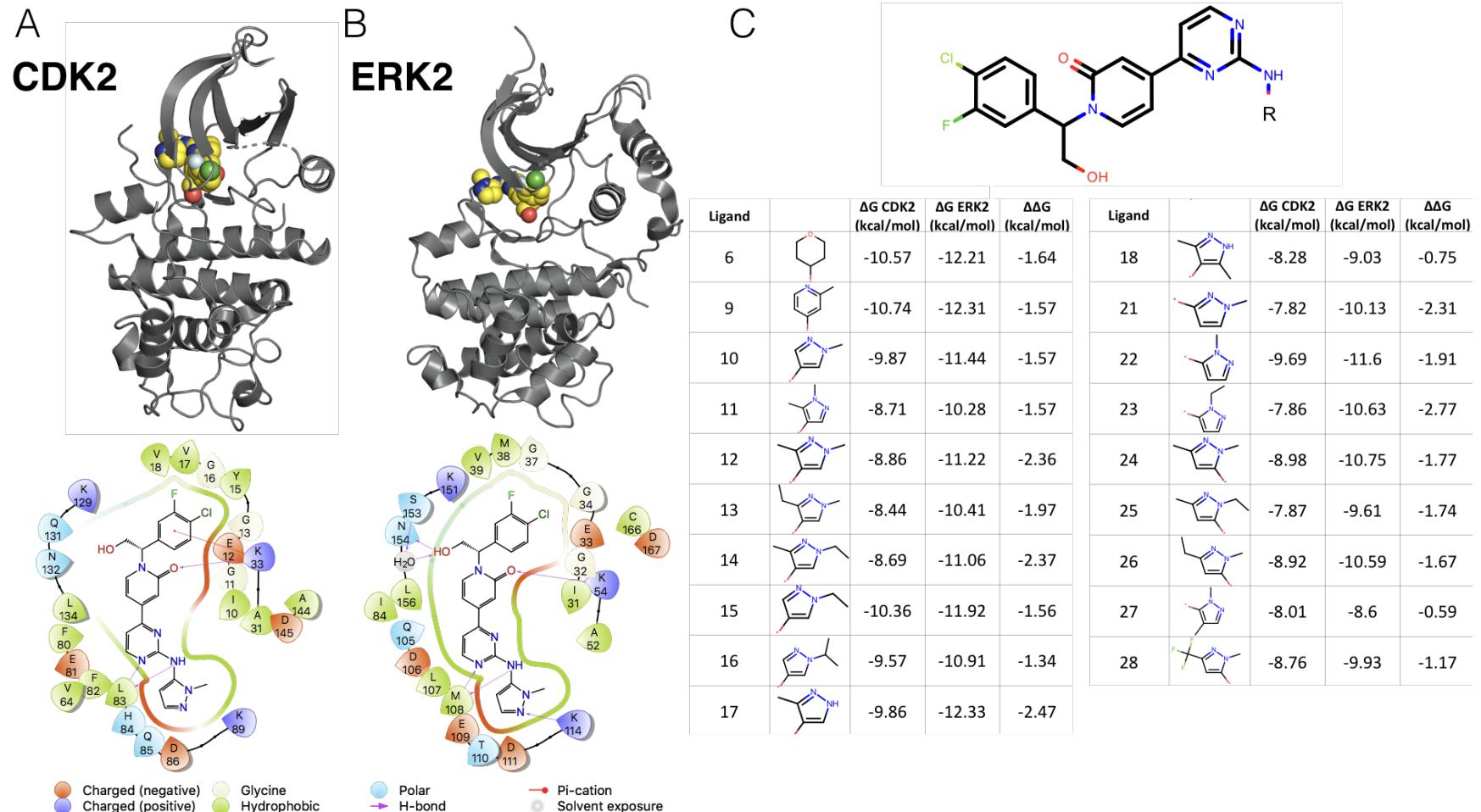


Figure 2.3: CDK2 and ERK2 selectivity dataset from Blake et al., 2016

(A) (top) Crystal structure of CDK2 (5K4J) shown in gray cartoon and ligand 22 shown in yellow spheres. (bot) 2D interaction map of ligand 22 in the binding pocket of CDK2 (B) (top) Crystal structure of ERK2 (5K4I) shown in gray cartoon with ligand 22 shown in yellow spheres. (bot) 2D interaction map of ligand 22 in the binding pocket of ERK2. (C) (top) Common scaffold for all of the ligands in the Blake dataset, with R denoting attachment side for substitutions. (bot) Table showing R group substitutions and experimentally measured binding affinities and selectivities. Ligand numbers correspond to those used in publication.

2.4.5 FEP+ calculations show accurate potency predictions for ERK2/CDK2 and larger errors for CDK2/CDK9

The FEP+ predictions of single target potencies (ΔG) showed good accuracy for the CDK2 and ERK2 dataset (Figure 2.4, top), with an RMSE of $0.37^{0.57}_{0.16}$ and $0.53^{0.81}_{0.22}$ kcal/mol, respectively. All of the CDK2 and ERK2 potencies were predicted within 1 log unit of the experimental value. Despite the high accuracy for the single target potency predictions, the selectivity ($\Delta\Delta G_{selectivity}$) predictions show an RMSE of $0.81^{1.26}_{0.37}$ kcal/mol, with all of the predictions falling within 1 log unit of the experimental values (Figure 2.4, top right panel).. Despite the high accuracy of the predictions, the narrow dynamic range and high uncertainty from experiment and calculation obscures any signal in the data. The CDK2 and CDK9 datasets show higher errors in the potency predictions, with an RMSE of $1.39^{2.05}_{0.58}$ and $1.71^{2.61}_{0.61}$ kcal/mol respectively. There are a number of outliers that fall outside of 1 log unit from the experimental value. While the higher per target errors make predicting potency more difficult, the selectivity predictions show a much lower RMSE of $0.74^{1.25}_{0.31}$ kcal/mol. This suggests that some correlation in the error is leading to fortuitous cancellation of systematic error, leading to more accurate than expected predictions of $\Delta\Delta G_{selectivity}$.

this will change
if we take of the
ff uncertainty
error bars

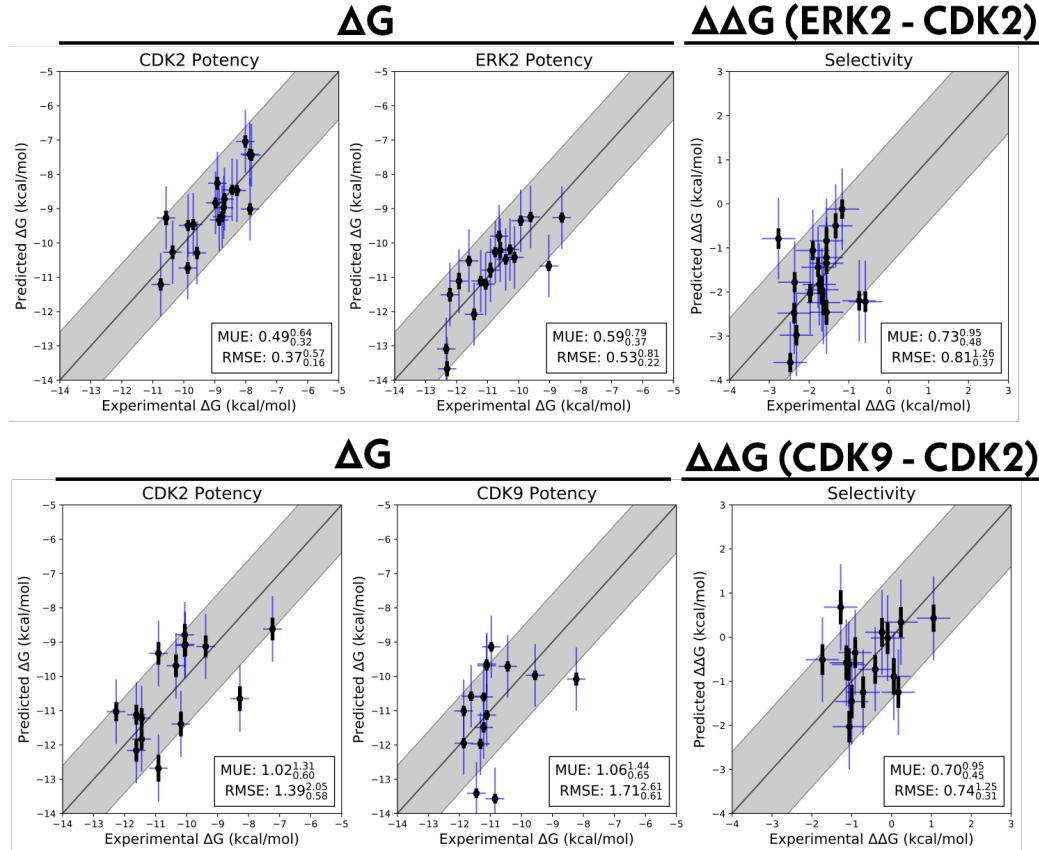


Figure 2.4: Relative free energy calculations can accurately predict potency, but show larger errors for selectivity predictions. Single target potencies and selectivities for CDK2/ERK2 from the Blake datasets (*top*), and CDK2/CDK9 (*bottom*) from the Shao datasets. The experimental values are shown on the X-axis and calculated values on the Y-axis. Each data point corresponds to a ligand for a given target. All values are shown in units of kcal/mol. The horizontal error bars show the assumed experimental uncertainty of 0.3 kcal/mol[55]. To better highlight outliers that are unlikely due simply to forcefield errors, we presume the forcefield error ($\sigma_{FF} \approx 0.9$ kcal mol⁻¹ [53]) also behaves as a random error. We show the total estimated statistical and forcefield error ($\sqrt{\sigma_{FF}^2 + \sigma_{calc}^2}$) as vertical blue error bars. The black vertical error bars correspond to the statistical error (σ_{calc}). The black line indicates agreement between calculation and experiment, while the gray shaded region represent 1.36 kcal/mol (or 1 log unit) error. The MUE and RMSE are shown on each plot with bootstrapped 95% confidence intervals.

2.4.6 Free energy calculation errors are correlated, accelerating selectivity optimization

To quantify the correlation coefficient (ρ) of the errors in our calculations, we built a Bayesian graphical model, as described in the methods section. Briefly, we modeled the absolute free energy (G) of each ligand in each phase (complex and solvent) as in equation 2.5. The model was chained to the FEP+ calculations by providing the $\Delta G_{phase,ij,target}^{calc}$ as observed data, as in equation 2.7. As in equation, the experimental data was modeled as a normal distribution centered around the true free energy of binding ($\Delta G_{i,target}^{true}$) corrupted by experimental error, which is assumed to be 0.3 kcal/mol from previous work done to quantify the uncertainty in publicly available data [55]. The reported IC50 values from each dataset were treated as data observations (Equation 2.10) and the $\Delta G_{i,target}^{true}$ was assigned a weak normal prior (Equation 2.11). The correlation coefficient was calculated for each sample according to equation 2.12. The correlation coefficient ρ for the CDK2/ERK2 calculations was quantified to be $0.5^{0.33}_{-0.23}$, indicating that the errors are largely uncorrelated between ERK2 and CDK2 (Figure 2.5A, right). The joint marginal distribution of the error (ϵ) for each target is symmetric, which is expected for cases in which ρ is 0 (Supp. Figure 2.7). Despite the weak correlation in errors, the high per target accuracy of these calculations should have a 2-3x speed up for 1 log unit selectivity optimization, and a 20-30x speed up for 2 log unit selectivity optimization (Figure 2.5A, right). The CDK2/CDK9 calculations show strong evidence of correlation, with a correlation coefficient of $0.70^{0.82}_{0.57}$ (Figure 2.5B, right). The joint marginal distribution of errors is strongly diagonal, which is expected based on the value for ρ (Figure 2.5B, left). The high correlation in errors leads to a speed up of 4-5 for 1 log unit selectivity optimization and 30-40x for 2 log unit selectivity optimization (Figure 2.5B, right), despite the much

higher per target errors. Quantifying ρ for these calculations enables estimation of $\sigma_{selectivity}$, which is useful for estimating expected error for prospective studies, where the experimental values for $\Delta\Delta G_{selectivity}$ are not yet known. Based on the distribution quantified for ρ , the expected $\sigma_{selectivity}$ for the CDK2/CDK9 calculations is between 0.76 and 1.16 kcal/mol (Supp. Figure 2.8), which is in good agreement with the bootstrapped RMSE (Figure 2.4, bottom). For the CDK2/ERK2 calculations, $\sigma_{selectivity}$ is expected to fall between 0.82 kcal/mol and 1.10 kcal/mol (Supp. Figure 2.8), which is also in good agreement with the bootstrapped RMSE (Figure 2.4, top).

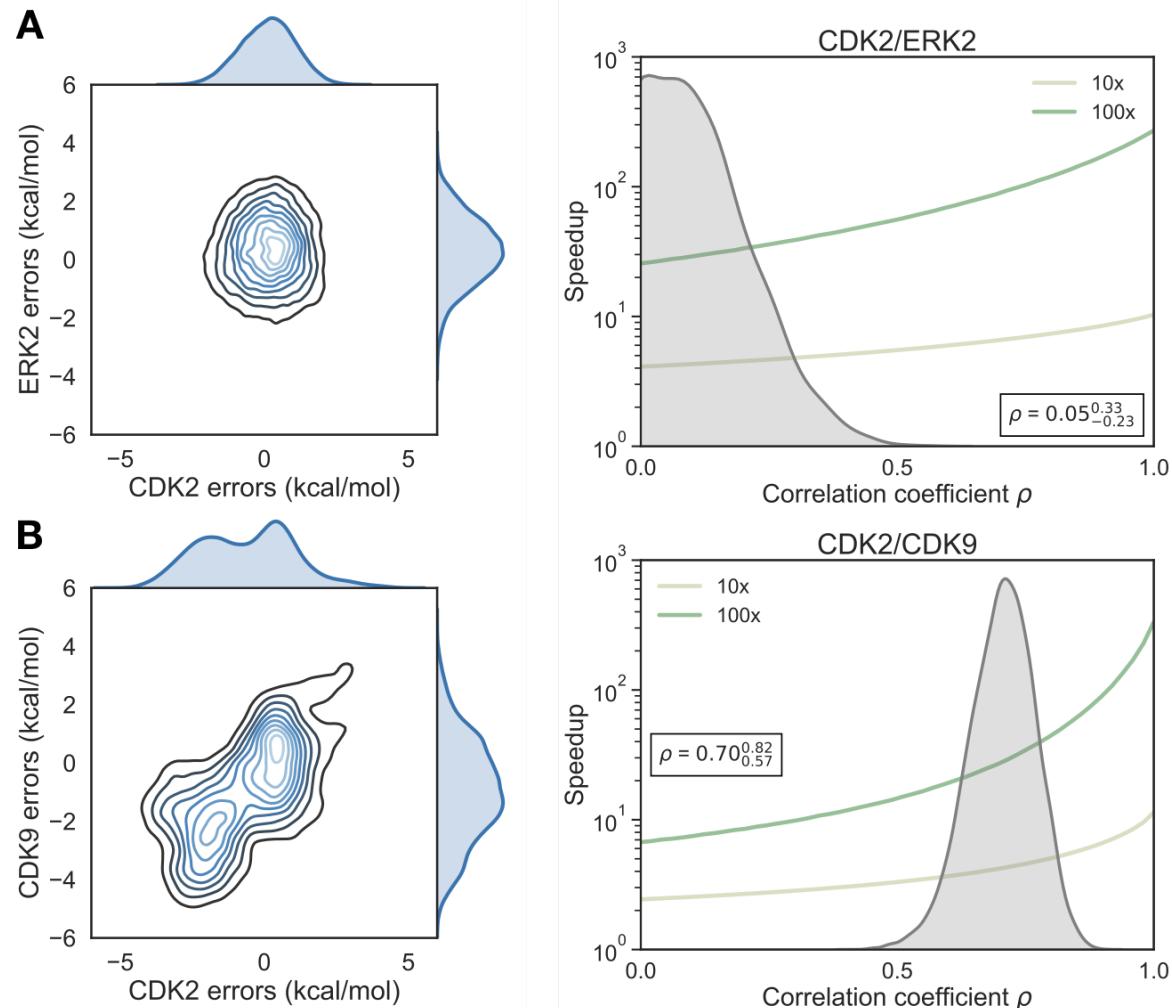


Figure 2.5: **Correlation in selectivity prediction errors can be used to accelerate selectivity optimization**

(A) (left) The joint posterior distribution of the prediction errors for CDK2 (X-axis) and ERK2 (Y-axis) from the Bayesian graphical model. (right) Speedup in selectivity optimization (Y-axis) as a function of correlation coefficient (X-axis). The posterior marginal distribution of the correlation coefficient (ρ) is shown in gray, while the expected speed up is shown for 100x (green curve) and 10x (yellow curve) selectivity optimization. The inserted box shows the mean and 95% confidence interval for the correlation coefficient. (B) (left) The same as above, with CDK2 (X-axis) and CDK9 (Y-axis). (right) As above, for the CDK2/CDK9 calculations.

2.5 Discussion and Conclusions

We have demonstrated, using a simple numerical model, the impact that free energy calculations with even weakly correlated errors can have on speeding up the optimization of selectivity in small molecule kinase inhibitors. While the expected speed up is dependent on the per target error of the method (σ_{target}), the speedup is also highly dependent on the correlation of errors made for both targets. Unsurprisingly, free energy methods have greater impact as the threshold for selectivity optimization goes from 10x to 100x. While 100x selectivity optimization is difficult to achieve, the expected benefit from free energy calculations is also quite high, with 1 and 2 order of magnitude speedups possible. To quantify the correlation of errors in two example systems, we gathered experimental data for two congeneric ligand series with experimental data for CDK2 and ERK2, as well as CDK2 and CDK9. These datasets, which had crystal structures for both targets with the same ligand co-crystallized, are exemplify the difficulty in predicting selectivity. The dynamic range of selectivity for both systems is incredibly narrow, with most of the perturbations not having a major impact on the overall selectivity achieved. Furhter, the data was reported with unreliable experimental uncertainties, which makes quantifying the errors made by the free energy calculations difficult. This issue is common when considering selectivity, as many kinase-oriented high throughput screens are carried out at a single concentration and not highly quantitative. Work is being done to increase the availability of publicly available, high quality biophysical data for selectivity measurements, which will benefit future work on predicting selectivity using physical models and machine learning techniques.

insert a discussion of outliers in the FEP+ calculations here?

discuss quantification of rho (is there a way to know a priori what rho might be?)

discussion of the expected speedup and sigma based on the quantification of rho

extensions include: separating statistical error (unless this gets added into the results section)

Do we want to mention protein mutation FEP+ to enable cycle closure analysis for certain highly related targets?

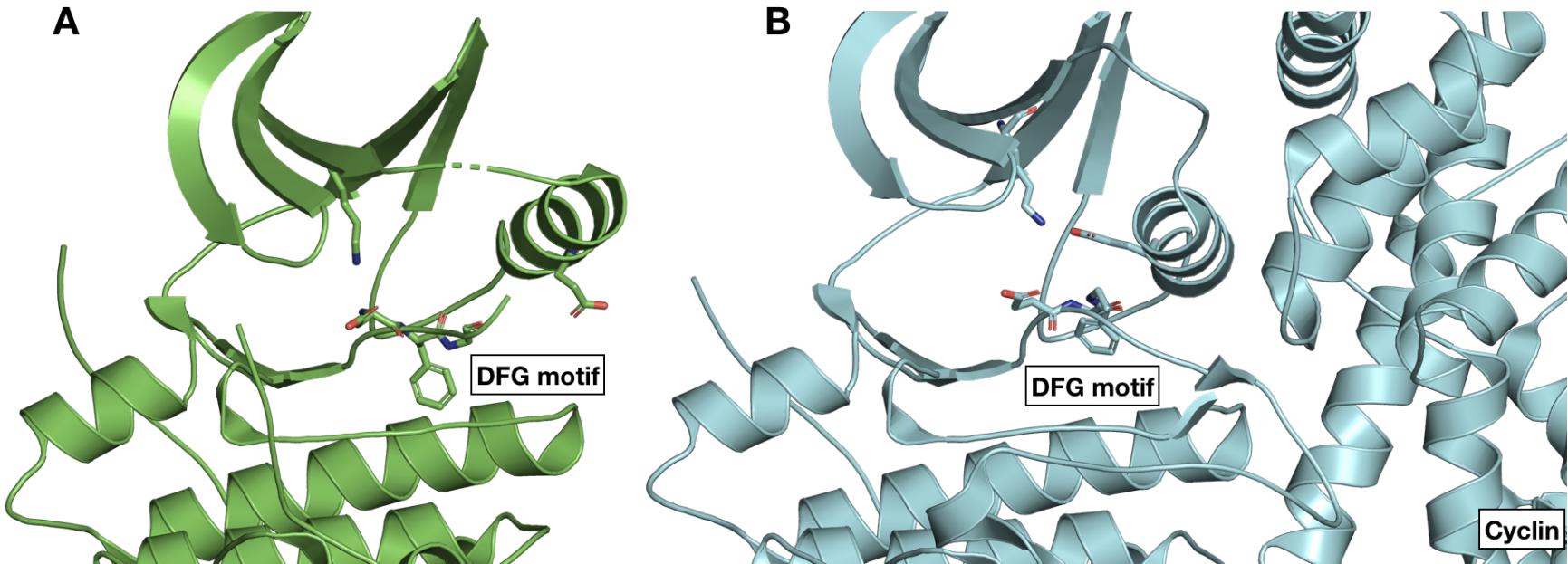


Figure 2.6: CDK2 adopts an inactive conformation in the crystal structure used for the CDK2/ERK2 calculations (A) CDK2 (5K4J) adopts an inactive conformation in the absence of its cyclin. The DFG motif is in a DFG-out conformation, with the α C helix rotated outwards, breaking the salt bridge between K33 and E51 (Uniprot numbering) that is typically a marker of an active conformation. Notably, the Phe in the DFG motif does not completely form the hydrophobic spine due to the rotation of the α C helix [105] (B) The CDK2 structure used for the CDK2/CDK9 calculations (4BCK) contains cyclin A and adopts a DFG-in/ α C helix-in conformation that forms the salt bridge between K33 and E51. This is typically indicative of a fully active kinase [18, 104].

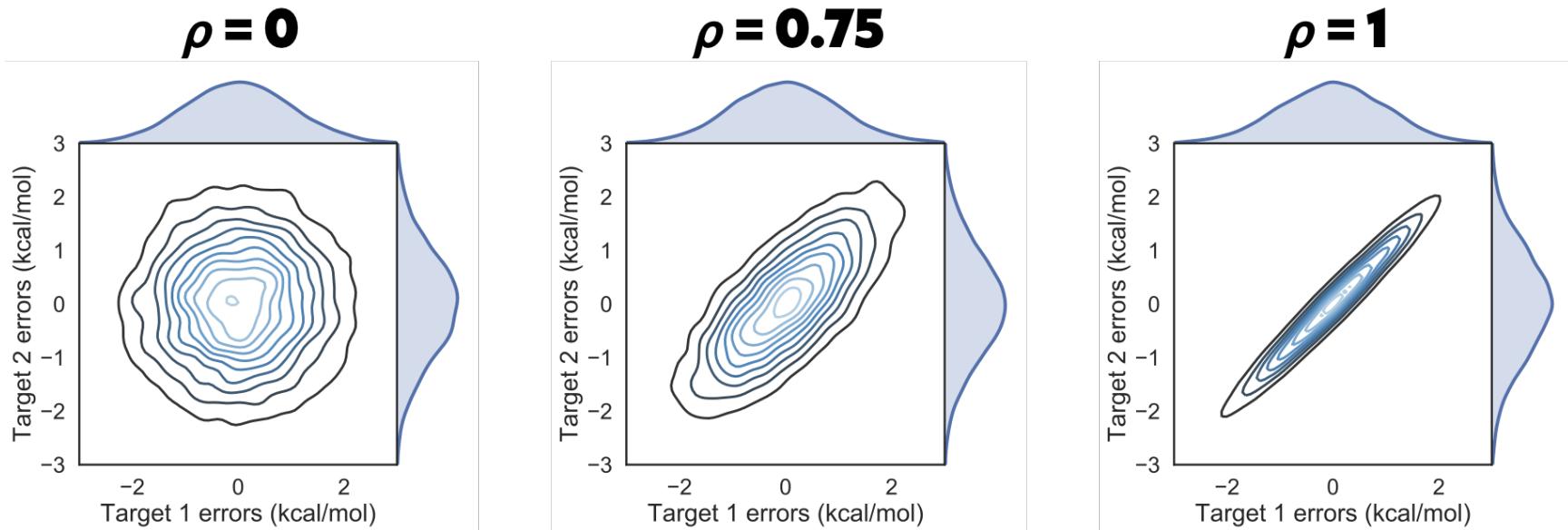


Figure 2.7: **Correlation coefficient ρ controls the shape of the joint marginal distribution of errors** As ρ increases, the joint marginal distribution of errors become more diagonal. Each panel shows 10000 samples drawn from a multivariate normal distribution centered around 0 kcal/mol, where the per target error was set to 1 kcal/mol and ρ to the value indicated in bold over the plot.

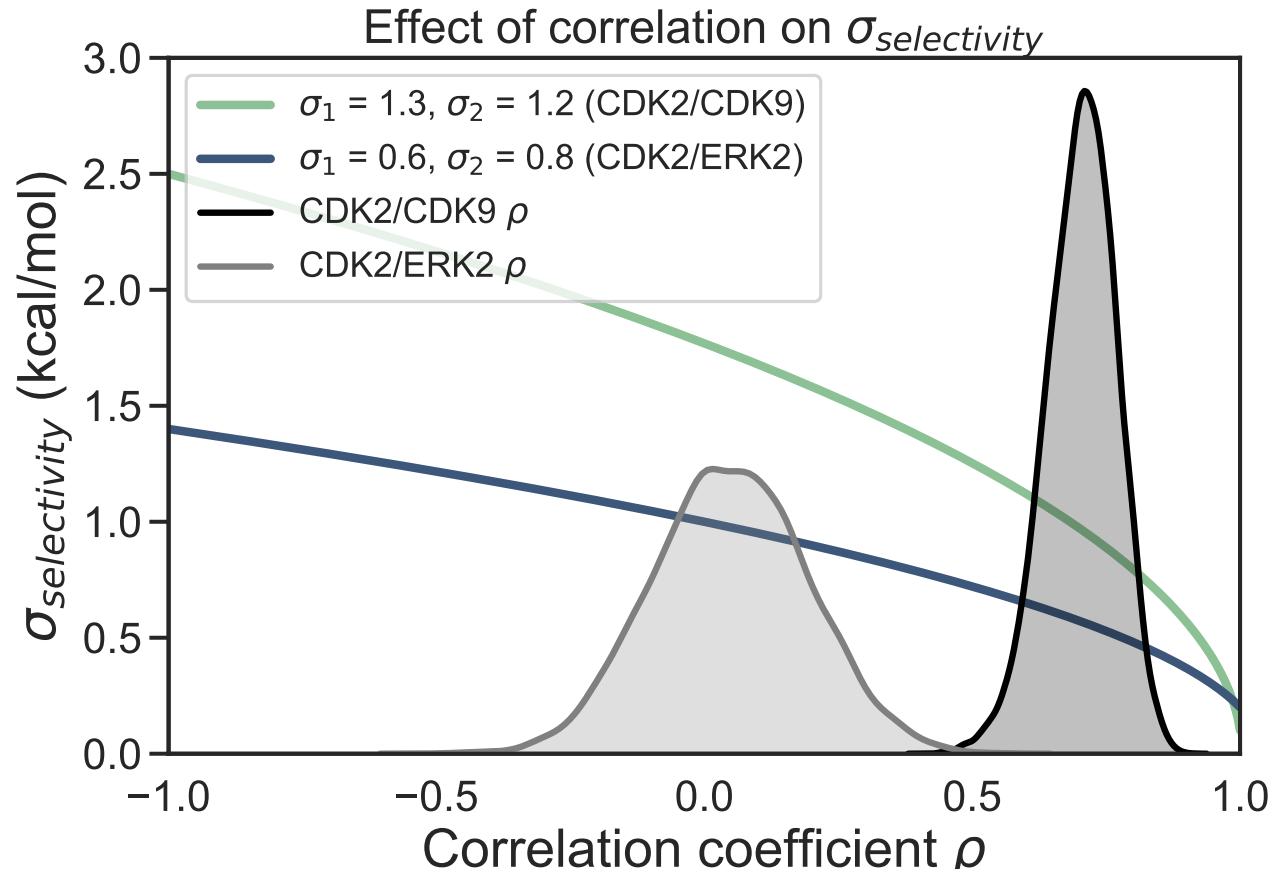


Figure 2.8: Correlation reduces the expected error for selectivity predictions As corelation coefficient ρ increases, $\sigma_{selectivity}$ decreases. The intersection between CDK2/CDK9 $\sigma_{selectivity}$ (green curve) and ρ (black distribution) indicates the range of expected $\sigma_{selectivity}$ values. The intersection for CDK2/ERK $\sigma_{selectivity}$ (blue curve) and ρ (gray distribution) suggests the expected $\sigma_{selectivity}$ range for that set of calculations.

CHAPTER 3

**UNDERSTANDING THE FUNCTIONAL IMPACT OF MTOR CLINICAL
KINASE MUTATIONS USING PHYSICAL MODELING**

3.1 Introduction

3.1.1 mTOR forms the catalytic core of protein complexes that control a number of cellular processes

mTOR (mammalian target of rapamycin) is a serine-threonine kinase that controls a number of cellular processes [106, 107] by integrating signaling from the MAPK and PI3K pathways [27]. mTOR forms the catalytic core of heteromeric protein complexes, mTORC1 and mTORC2, that differ in the regulatory proteins that decorate the kinase. mTORC1 is defined by RAPTOR [108, 109] and PRAS40 [110], while mTORC2 is characterized by RICTOR [111]. Each complex contains a number of shared regulatory proteins, such as mLST8 [112] and DEPTOR [113]. mTOR itself is an atypical kinase, and includes a number of insertions that deviate from the canonical kinase fold. The C-terminal fragment crystal structure (Figure 3.1) shows that the kinase domain is hugged by a FAT domain [114], forming a C-shaped clamp around the kinase domain. The FAT domain forms a number of regulatory salt bridges that have been implicated in controlling the activity of the kinase [114]. The FK506-rapamycin-binding (FRB) domain hangs over the opening cleft to the active site [114] (Figure 3.1). The FRB domain is exposed in mTORC1, while it is thought to be occluded and inaccessible to FKBP12 in mTORC2 [115]. Complex activity is hypothesized to be regulated via restricted access to the active site by architectural elements as well as

rapalog-mediated recruitment of FKBP to the FRB domain [116].

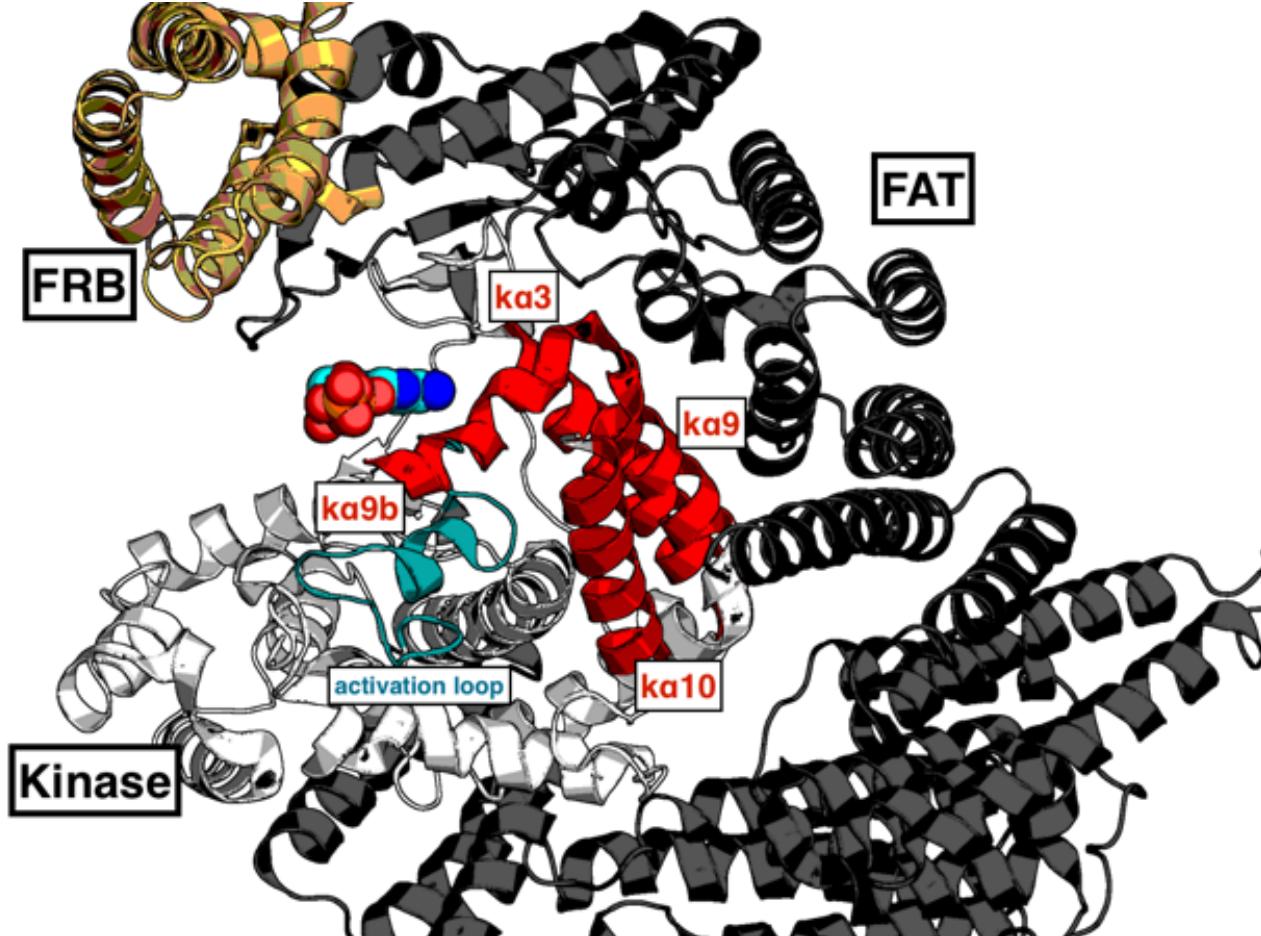


Figure 3.1: **mTOR is an atypical kinase with a number of regulatory domains** mTOR (PDBID: 4JSV) is shown as a cartoon diagram. The kinase domain (white) has a number of structural features highlighted, such as the activation loop (teal) as well as a network of regulatory α helices (red) in the active site. The FAT domain (black) clamps around the kinase domain and forms a number of salt bridge and hydrophobic contacts with the kinase domain. The FRB domain (gold) hangs over the active site clef and is the site of rapalog binding and rapalog-mediated FBKBP recruitment. This figure reprinted with permission of James Hsieh

mTORC1 controls a number of biological processes by integrating multiple upstream signals (Figure 3.2). Tuberous sclerosis complex 1/2 (TCS1/2) integrates growth factor signaling from the PI3K/AKT pathway [117, 118] and DNA damage from AMPK [119]. TSC1/2 acts as GTPase activating protein (GAP) for RHEB [120], which binds to and activates mTORC1 at the lysosomal surface. mTORC1 localization is controlled by the second major signal it integrates: amino acid availability. Upon stimulation by amino acids, the Ragulator [121] recruits inactive mTORC1 from the cytosol to the lysosomal surface [122–124], enabling interaction with RHEB. Once active, mTORC1 controls cell growth and protein synthesis by phosphorylating and activating S6K at T389 and inactivating 4EBP1 via a phosphorylation at S65 [106, 125]. mTORC1 also activates SREBP1/2, which regulates lipid biosynthesis [126]. mTORC1 activation inhibits autophagy through via ULK1. Recent work suggests that mTORC1 can also control the biophysical properties of the cytoplasm by regulating crowding [127], impacting the rate of diffusion and expanding the already diverse array of essential processes that mTOR controls.

3.1.2 mTOR is targeted by inhibitors with two distinct mechanisms of action

Due to the central role mTOR plays in a wide array of biological processes, it has emerged as the target of extensive drug discovery programs. There are two distinct classes of inhibitors developed to target mTOR: ATP-competitive inhibitors and allosteric inhibitors [128, 129]. Temsirolimus [130] and Evirolimus [131] are FDA-approved inhibitors [3] that are thought to inhibit mTOR through recruiting FKBP family members, such as FKBP12, to the FRB

domain [132]. Evidence suggests FKBP12 recruitment may inhibit mTORC1 activity by inducing a conformational change that prevents S6K binding [133]. Further incubation leads to destabilization and disassembly of the protein complex, which is consistent with the time-dependent nature of rapalog inhibition of mTORC1 mediated 4EBP1 phosphorylation [133]. The rapalogs are potent inhibitors of mTORC1, while mTORC2 requires chronic rapalog treatment and is highly dependent on FKBP expression levels [134]. A cryo-EM structure of TORC2 suggests that the FRB domain is inaccessible in TORC2, explaining the relative insensitivity of mTORC2 to rapalog inhibition [115]. Rapalogs have also been shown to only partially inhibit mTORC1, often failing to reduce 4EBP phosphorylation levels [107], despite efficacy at preventing S6K phosphorylation. There are an array of ATP-competitive inhibitors developed to target mTOR and other members of the PI3K family of kinases, such as MLN0128 [135, 136], INK-228 (TAK-128) [137], AZD8055 [138], PKI-587 [139], and BEZ-235 [140]. ATP-competitive inhibitors exhibit a wide range of selectivity, with many inhibiting mTORC1, mTORC2, and multiple isoforms of PI3K. ATP competitive inhibitors target the catalytic activity of mTOR and more fully inhibit mTORC1 and mTORC2 phosphorylation of downstream targets [107]. While many of these inhibitors have struggled clinically due to modest clinical benefits and high toxicity leading to low tolerability [141], there is considerable excitement about developing mTOR inhibitors to treat a number of diseases. ATP-competitive inhibitor BEZ-235 has shown promising results in preclinical models of lapatinib-resistant PI3K hyperactivated breast cancer [142]. Dual targeted catalytic inhibitors that target both PI3K and mTOR have shown success by abrogating reactivation of Akt, which is caused by the alleviation of negative feedback due to long term treatment with mTOR inhibitors [143]. For example, a dual-inhibitor of mTORC and PI3K α , PI-103, showed promising activity against

glioma xenografts [144].

mTOR inhibitors show the most promise in treating diseases with mTOR pathway alterations pathway [145]. Despite initial success, resistance mutations in the FRB or kinase domain have been observed [146]. Work has also been done to combine the two mechanisms of action in Rapalink-1, which tethers rapamycin and MLN0128 [147] with a flexible linker and overcome such acquired resistance. Rapalink also shows improvements over ATP-competitive inhibitors or rapalogs outside of the context of resistance. In glioblastoma models, ATP-competitive inhibitor MLN0128 and allosteric inhibitor rapamycin showed poor *in vivo* efficacy. In contrast, RapaLink-1, by overcoming poor residence time and potency, was able to durably inhibit mTORC1 and cross the blood-brain barrier [148].

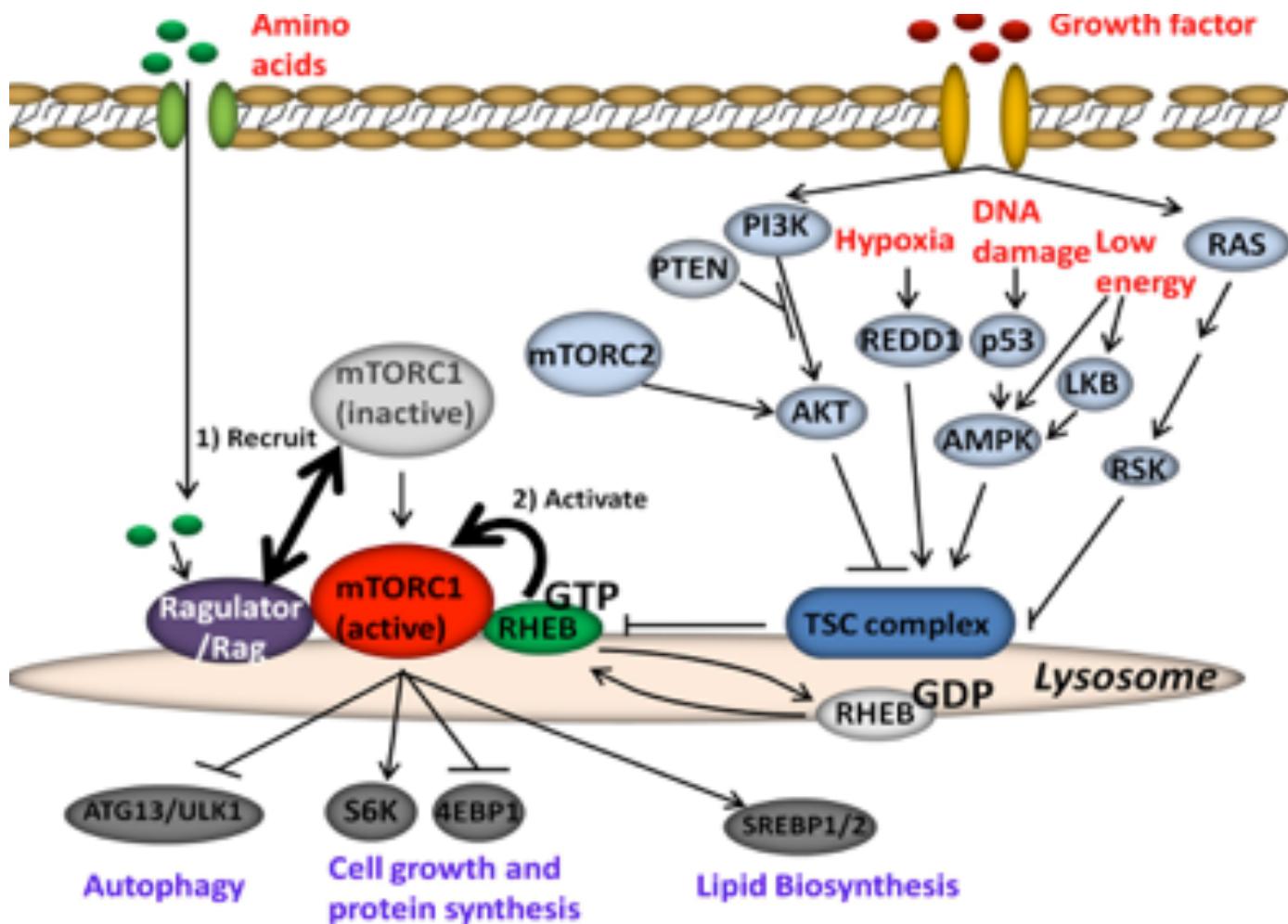


Figure 3.2: **mTORC1 integrates signaling from a number of inputs** The signaling pathway of mTOR involves integrating signaling from the growth factors, amino acids, hypoxia, and DNA damage. Integrating such signals controls the localization and activity of mTORC1, which controls autophagy, cell growth, and macromolecule synthesis through phosphorylation of downstream targets. **This figure courtesy of James Hsieh**

3.1.3 mTOR signaling is dysregulated in cancer by hyperactivating missense mutations

mTOR pathway alterations have been observed in a wide array of cancer types [149] and extensively characterized. Less well studied are missense mutations in mTOR itself. An exceptional responder in a phase 1 clinical trial of pazopanib and everolimus with metastatic urothelial carcinoma lead to the identification of two missense mutations in mTOR [145], E2014K and E2419K. Subsequent work identified 33 MTOR mutations gathered from publicaly available tumor sequencing data, and found that a number of them activated the mTOR pathway and conferred sensitivity to mTOR inhibitors when engineered into various cancer cell lines *in vitro* [150]. While these missense mutations are pulled from an array of different cancer types, mTOR missense mutations have been observed in about 15% of clear cell renal cell carcinoma [151]. Characterization of these mutations revealed that not only do many of these mutations hyperactivate mTOR (Figure 3.3), but they can be grouped into complementation groups by determining which double mutations cause hyperactivation at levels higher than the constituent single mutants alone [152]. This suggests that these mutations can activate mTOR through different, seemingly complementary mechanisms.

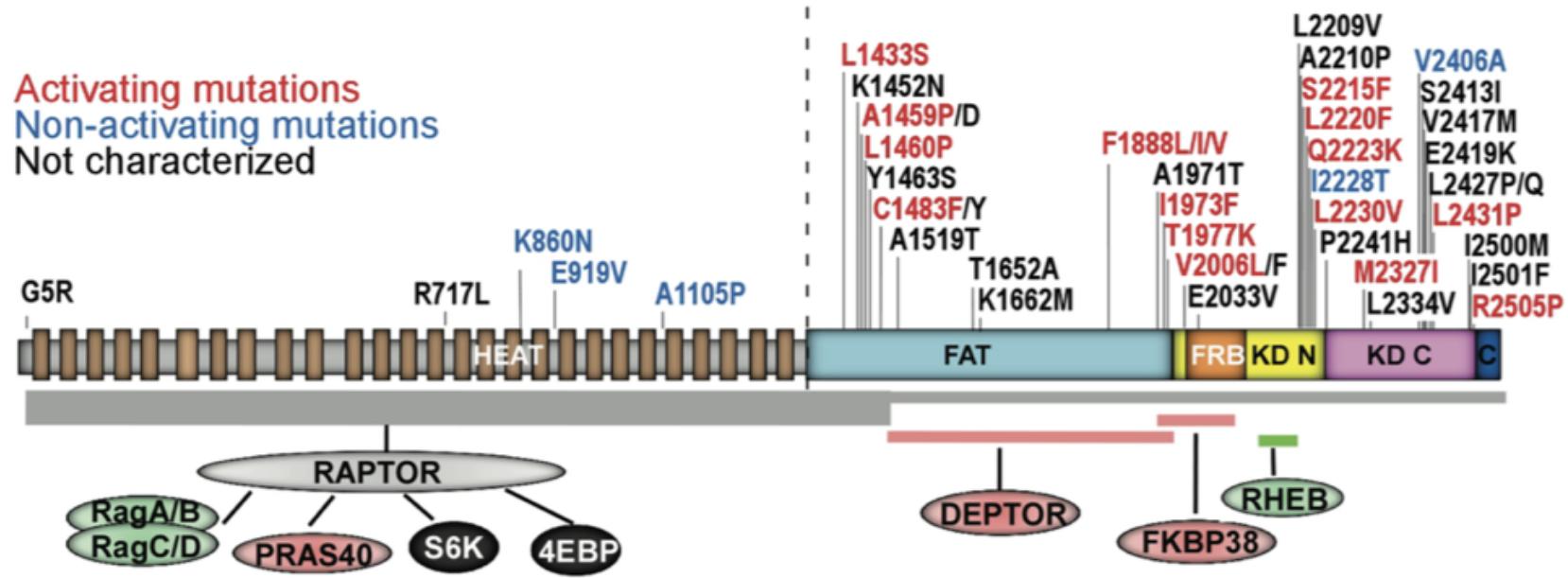


Figure 3.3: Hyperactivating mTOR missense mutations have been observed in cancer Diagram shows the domain structure of mTOR, its regulatory interaction partners (negative regulators in pink, positive regulators in green, and a dual-role regulator in gray), and the substrates of mTORC1 complex. The positions within mTOR that are involved in the interaction with the regulatory partners are highlighted below the domain structure. The thickness of the horizontal bar of RAPTOR-mTOR interaction indicates the relative binding affinity. mTOR missense mutations derived from ccRCC are mapped and color coded to summarize their respective effects on mTORC1 signaling (activating mutations in red). KD N, kinase domain N lobe; KD C, kinase domain C lobe. **This figure reprinted with permission of James Hsieh and the Journal of Clinical investigation [152].**

3.1.4 Using physical modeling to understand the functional impact of mTOR mutations

To begin to understand the impact of these missense mutations at an atomistic level, we performed massively parallel molecular dynamics [153] using the computing resource Folding@Home [154]. Molecular Dynamics simulations have been used previously to understand the mechanism of oncogenic and resistance mutations on the structure and activation of EGFR [64, 65]. They have also been applied to understanding missense mutations in p53 [67], CLIC2 [68], opsin [69], and a host of oncogenes and tumor suppressors[155]. Using the previously solved crystal structure [114], we built models for 45 single, and 190 double mutations, far more than could be crystallized individually [152]. We analyzed the simulations for changes in contact formation, as well as changes in order parameters for activation mined from other kinase studies and previous work on mTOR biochemistry. We also piloted alchemical free energy calculations [156] on a number clinically observed mutants, to compute physical, testable properties such as change in affinity for ATP-competitive inhibitors and ATP itself. In doing so, we identify promising candidates for resistance mutations to an ATP competitive inhibitor, and lay the ground work for future studies on the application of these methods to studying the functional impact of mutations on inhibitor binding. Taken together, this work forms the beginning of a comprehensive analysis of the impact of these mutations on structure and small molecule binding, which has implications for the treatment of patients with these mutations.

3.2 Methods

3.2.1 Molecular dynamics simulations

This work was performed and previously described in reference [152]. The canonical wild-type mTOR sequence for the UniProt-annotated PI3K/PI4K domain span (residues 2182-2516) was modeled onto the X-ray structure of mTOR from chain A of RCSB entry 4JSN using the Ensembler automated simulation setup tool [157] with default parameters. A second set of simulations was performed using the full length, C-terminal fragment of the mTOR crystal structure (Uniprot sequence residues 1376-2549) from RCSB entry 4JSN, hereafter called the mTOR kinase+FAT simulations. This sequence was modeled onto the crystal structure using Ensembler with default parameters.

All residues were assigned default protonation states typical of pH 7.4. The AMBER 99SB-ILDN [158] forcefield was used for the protein along with the TIP3P solvent model [159] with neutralizing monovalent Na+ or Cl- counterions. The resulting simulation box had 80,983 atoms. The OpenMM 6.2 simulation package [160] was used for all minimization, equilibration, and production simulations. Equilibration simulations utilized Langevin dynamics with a timestep of 2 fs and collision rate of 20/ps, along with a Monte Carlo barostat with molecular scaling and update interval of 50 steps, with temperature and pressure control set to 300 K and 1 atm. Particle-mesh Ewald (PME) with default parameters was used for long-range electrostatic treatment, direct-space and Lennard-Jones interactions were truncated at 9 Å, and a long-range dispersion correction was employed. Bonds to hydrogen were constrained using CCMA [161] using the default tolerance of 1e-5, and waters were rigidly

constrained using SETTLE [162]. The Ensembler package [157] handles energy minimization and refinement in implicit solvent followed by a short minimization and equilibration step in explicit solvent prior to production simulations.

Production simulations of the wild-type kinase domain and full length C-terminal fragment were run on Folding@home [154] using a simulation core based on OpenMM 6.2 and the same simulation parameters, with the exception of a reduced collision rate of 1/ps. The structure obtained after 589.5 ns—which had relaxed much of the initial loop-modeling-induced structural artifacts—was used as a starting model for further modeling of mutations and subsequent production simulations. To model mTOR mutants and wild-type behavior, PDBFixer v1.2 [163], part of the Omnia molecular simulation suite, was used to generate mutant versions of the mTOR kinase domain and full length C-terminal fragment using the relaxed wild-type structure. Subsequent simulation steps utilized reaction-field electrostatics with a cutoff of 10A in place of PME to allow longer trajectories to be generated. Proteins were resolved in TIP3P water with NaCl counterions to neutralize the system and produce an environment of approximately 150 mM NaCl to using a padding of 11A around the kinase, resulting in systems of approximately 81K atoms. Langevin dynamics with a collision rate of 5/ps was used for subsequent production simulations of mutant and wild-type kinase domains, which also employed a simulation core based on OpenMM 6.2 on Folding@home. Simulation boxes were energy minimized with the OpenMM LocalEnergyMinimizer facility before subjecting them to subsequent dynamics. 20 replicate simulations of each mutant simulation box were run on Folding@home, with each replicate receiving a unique random number seed ensuring rapid decorrelation of trajectories. Each of the trajectories were 501 nanoseconds of simulation. The initial 100 ns of each simulation were discarded and subsequent simulation data was analyzed for structural

alternations indicative of rapid mutation- induced conformational changes.

3.2.2 Contact Map Analysis

Conformational changes were detected by generating a contact map, which shows the net change in the probability of forming a contact between a pair of residues from wild-type to mutant. To calculate these probabilities, mdtraj [164] was first used to calculate the distance between every residue pair based on the closest heavy atoms in each frame of the simulations after 100 ns. Using 5 angstroms as the threshold at which a contact was formed, the number of frames in which a contact was formed was divided by the total number of frames for each simulation, giving a probability for each residue pair in that simulation which could be averaged over the number of replicates per mutant. The wild-type probability was subtracted from the mutant, giving a net change in probability to form a contact for each residue pair in the protein. The structural images were generated using PyMOL to visually inspect areas of interest identified by the contact map over the course of each simulation.

3.2.3 Mean Contact formation over time

The fractional contact formation was calculated for 20 replicate 500ns trajectories. Analysis was carried out using 20ns sliding window chunks, with a single frame step (2fs) between each window. Contact cutoff at 4Å for closest heavy atom. Plotted is the mean \pm SEM, calculated as in Equation 3.2

$$\sigma = \sqrt{\frac{\sum^n(X_i - \mu_x)^2}{n - 1}} \quad (3.1)$$

$$SEM = \frac{\sigma}{\sqrt{n}} \quad (3.2)$$

Where σ in Equation 3.1 is the standard deviation, X_i fractional contact for a frame i , μ is the mean fractional contact for a 20ns window, and n is the total number of frames in that sliding window.

3.2.4 Alchemical free energy calculations

Structure and Ligand Preparation

Structures for mTOR (4JSP [114] (ATP) and 4JSX [114] (AZD8055)) were downloaded from the PDB [95]. Models were prepared from chain A of each structure using Schrödinger's PrepWizard (2016-1) [96], keeping only the canonical wild-type mTOR sequence for the UniProt-annotated PI3K/PI4K domain span (residues 2182-2516). The FAT and FRB domains were removed from the structure. All other chains were deleted. PrepWizard was used to add in hydrogens at pH 7.4 for both protein residues and the cocrystallized ligand. The protonation state of the cocrystallized ligand was assigned the lowest energy state using Epik at pH 7.4 ± 2 . Hydrogen bonding was optimized using PROPKA at pH 7.4 ± 2 . Each of the structures was minimized using OPLS3 [53] and an RMSD convergence cutoff of 0.3Å. The missing loops, due to their large size, were not modeled in, and were capped by PrepWizard.

ATP and AZD8055 were prepared for docking using Schrödinger's LigPrep

(2016-1). 3D structures were generated using OPL3, and ionization state determined by Epik at pH 7.4 ± 2 . All other settings were left on default. The lowest Epik state penalty state was selected for each ligand.

Docking

ATP and AZD8055 were docked into 4JSP and 4JSX, respectively, using Schrödinger's GLIDE (2016-1) [165–167]. The receptor grid was generated centered on the cocrystallized ligand using the default settings of the Receptor Grid Generation panel. None of the rotatable groups were allowed to rotate to improve computational efficiency. The ligands were docked into these receptor grids using the extra precision (XP) protocol. Ligand sampling was set to flexible, allowing for nitrogen inversions and sampling of different ring conformations. Epik state penalties were added to the docking score, although only one input for each ligand was used. A post docking minimization was performed on the top 5 poses for each ligand. The docking protocol was set to write out only the best pose for each ligand, on the basis of the docking score after minimization.

Protein mutation FEP+

Protein mutation FEP+ [75, 103, 168] was used to calculate $\Delta\Delta G_{mutation}$ for both ATP and AZD8055. The models generated above were parameterized using the default OPLS3 forcefield [97] that shipped with Maestro 2016-1. SPC parameters were used for the water model [169]. The FEP+ panel offers an automated workflow which only requires an input structure and specified mutations. The protocol carried out as described in reference [103], with only single replicates performed. The reported uncertainties in each $\Delta\Delta G_{mutation}$ are the BAR uncertainty

estimates [170, 171]. The calculations were run for the default 5 ns length.

3.3 Results

3.3.1 Missense mutations perturb the structure of mTOR kinase domain

To begin understanding the types of rearrangements missense mutations can induce in the local structure of mTOR, we employed contact map analysis to identify regions in which we detected changes in the formation of contacts from the wild type to mutant simulations. This analysis is described in depth in the methods section. Briefly, we first calculated the distance between every residue pair based on the closest heavy atoms in each frame of the simulations after 100 ns. For each residue pair, we calculated the probability of forming a contact across all of the simulation data and replicas, using 5 Å as the threshold at which a contact was formed. The wild-type probability was subtracted from the mutant, giving a net change in probability to form a contact for each residue pair in the protein. This analysis was performed only for the kinase domain simulations, as the full contact map for the full length C terminal fragment was too computationally intensive to compute for the 190 mutant simulations. Using these maps, it is possible to identify regions in which the mutant is causing a conformational change, both locally near the mutant and more distantly through some allosteric mechanism. As an example of the 45 different mutant kinase domain simulations, we show the kinase domain simulations of S2215F, a highly activating and recurrent missense mutation [?]. S2215F shows evidence of a

rapid, mutation-induced local rearrangement, causing disruption of the α -helix in which it occurs (Figure 3.4, region 1). This unfolding is evident in many of the replicates, shown in red in the region 1 panel highlighted by Figure 3.4. S2215F, shown in yellow in Figure 3.4, also causes unwinding and loosening of α -helix $\kappa\alpha 8$, despite this secondary structural element being far from where the mutant occurs (Figure 3.4, region 2).

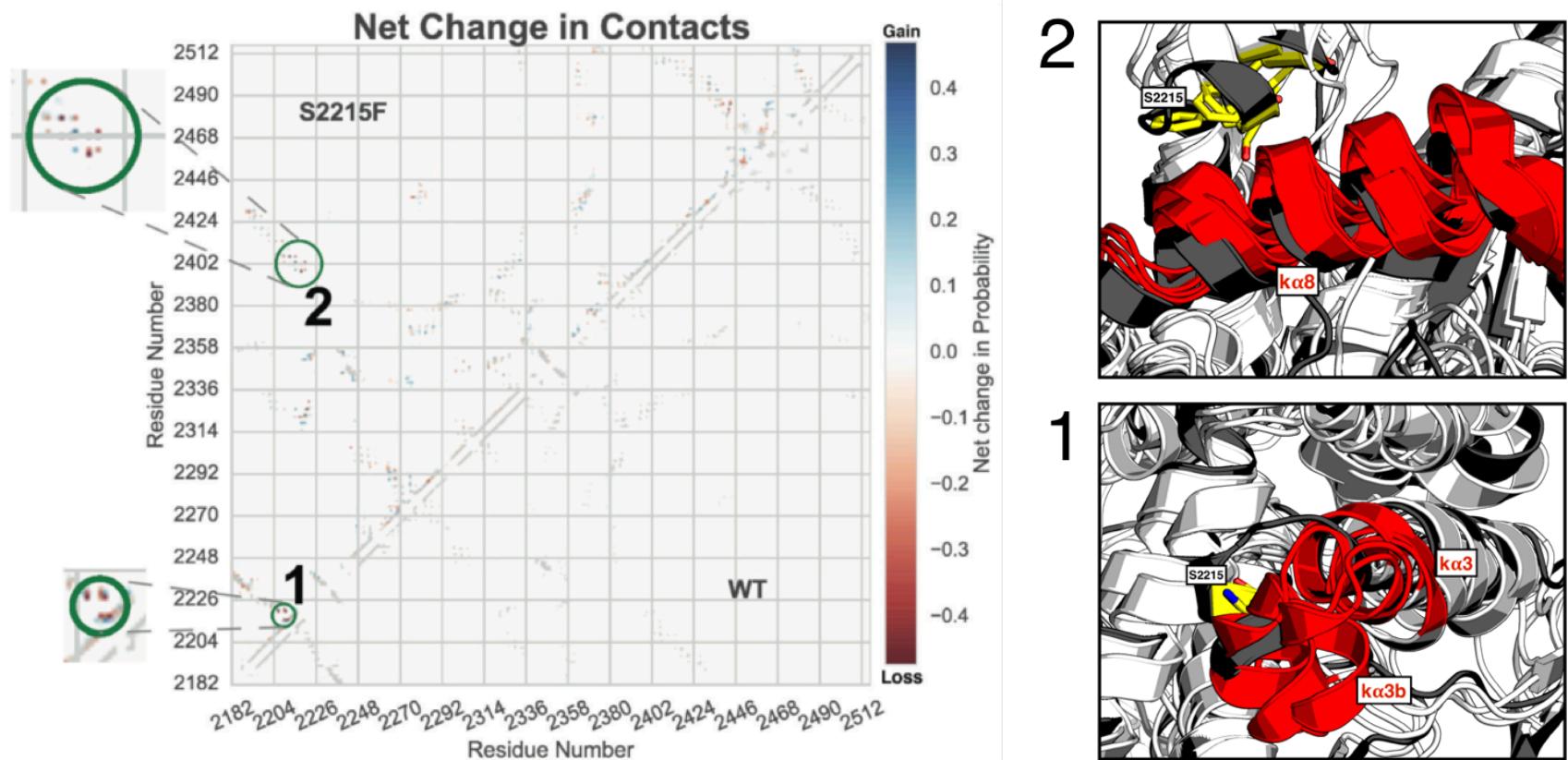
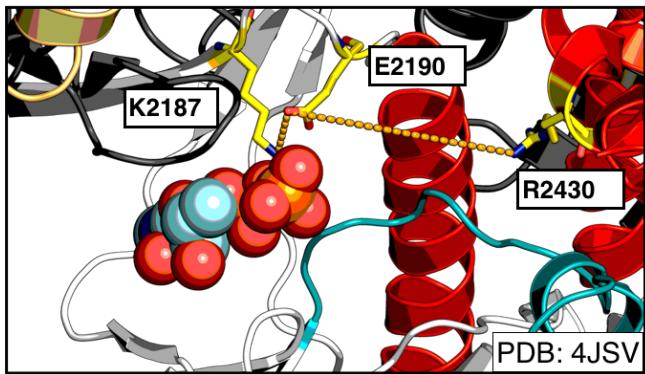


Figure 3.4: **Missense mutations can perturb local structure** Left Contact map showing the difference in probability of forming a contact between WT and mutant S2215F for the kinase domain simulations. Right Regions one and two highlighted in contact map, showing a structural perturbation in indicated helices. Starting structure is shown in gray, the residues indicated in the contact map are shown in red and residue 2215 is shown in yellow. All trajectories started from PDB: 4JSN. This left panel of this figure is a modified version of a figure that appears in [152]. Reprinted with permission of James Hsieh and the Journal of Clinical Investigation

3.3.2 Missense mutations do not appear to shift the formation of an active kinase domain

Despite the promise of being able to identify structural rearrangements in an automated, high-throughput fashion, contact map analysis is difficult to use to explain mechanistically how these mutants are activating. Seeking a more mechanistic understanding of how these mutants are activating, we looked at a common order parameter for activation from previous work on kinases [172]. In most kinases, there is a highly conserved lysine residue that coordinates ATP when it is bound in the active site. In mTOR, this lysine has been identified as K2187 [114]. To test whether mutations are activating mTOR by shifting the kinase domain into a more active conformation, we looked at the contact formation over time between residues K2187-E2190 (indicative of an active kinase), shown in Figure 3.5 in red, and E2190-R2430, shown in Figure 3.5 in blue. We hypothesized that formation of the E2190-R2430 contact suggests an inactive kinase domain, as the α C would need to rotate out and away from the active site, which has been observed in inactive kinases in previous work (Figure 3.5, inset panel) [172, 173]. Show on the right in Figure 3.5 is a representative panel of the activating mutants from the kinase domain simulations, as well as the control wild type. While some of the mutants seem less stable than the wild type, none of them appear to shift the distribution of the kinase domain further towards formation of the K2187-E2190 contact, indicative of activation. This is likely because the mTOR kinase domain adopts an active conformation in all of the available crystal structures, and the simulations largely stay in the local minima of the crystal structure. This is further supported by the hypothesis that the mTOR kinase domain is constitutively active, and is regulated by restriction of substrate access [106, 107, 114].



Indicative of active kinase
Indicative of inactive kinase

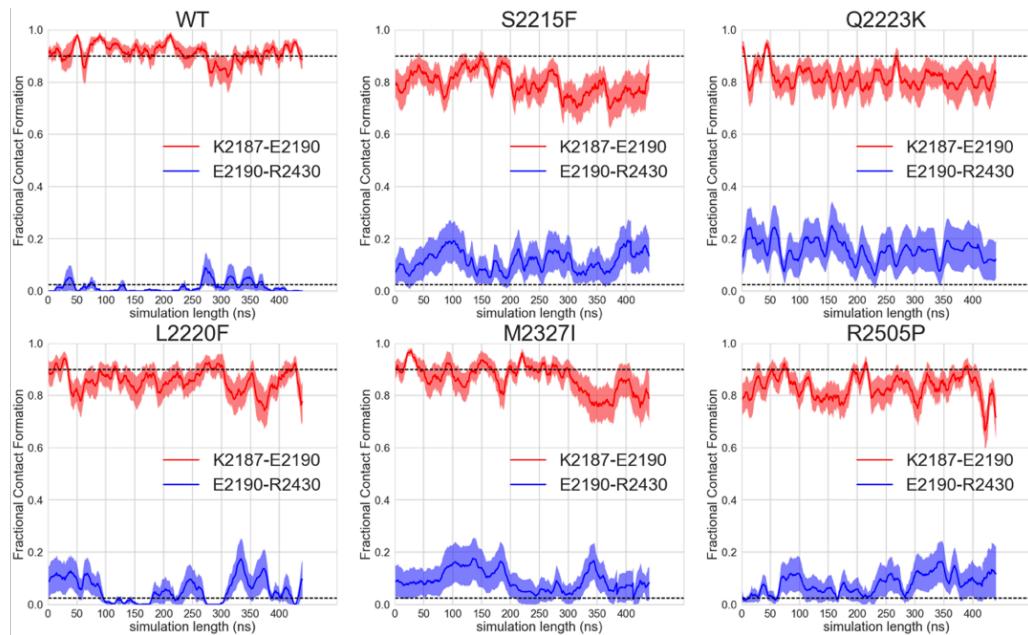


Figure 3.5: Missense mutations do not shift the population of the active conformation on a common activation order parameter
Right Fractional contact formation analysis for 20 500ns trajectories, analyzed in 20ns sliding window chunks. Contacts cutoff at 4 for closest heavy atom. Plotted is the mean \pm SEM. The dashed lines represent the proportion roughly populated in the WT simulations.
Left Illustration of the bond formed between K2187 and E2190 (red) or E2190 and R2430 (blue). ATP is shown in spheres. The kinase domain (white) is shown with the activation loop (teal) and the α helices (red). The FRB domain is shown in gold.

3.3.3 Missense mutations do not disrupt the formation of inhibitory salt bridges between the kinase and FAT domains

Another hypothesis for how the missense mutations may be hyperactivating looked at the role the FAT domain, which clamps around the kinase domain, plays in regulating the activity of mTOR. In previous work, point mutations that disrupted the formation of the salt bridge between E2419 and R1905 (Figure 3.6, red) activated TOR signaling in yeast and mammalian cell lines [174]. From this observation, we hypothesized that stable contacts formed between the kinase domain and FAT domain have a negative regulatory role on the activity of mTOR, and that disruption of these bonds through some allosteric mechanism would allow an oncogenic mutant to activate mTOR. Another set of contacts, Gln1941 to Gln2200 (Figure 3.6, blue) is conserved across many mTOR orthologs [114], suggesting a potential regulatory role. We also looked at several other salt bridges in the vicinity of these contacts: E1147 and K2218 (cyan), Q1425 and R2322 (magenta), and E1427 and R2322 (yellow). Shown in Figure 3.6 is a representative selection of mutations from the kinase domain, demonstrating that none of the salt bridges showed a prominent change from the wild type to mutant simulations. Of note, R2505P appears to have a disruption in the formation of the Q1941 and Q2200, E1147 and K2218, and E2419 and R1905 contacts, indicating that this may be a possible mechanism of activation for this mutant. However, this is difficult to say with certainty without functional experimental data or additional computation to confirm this result.

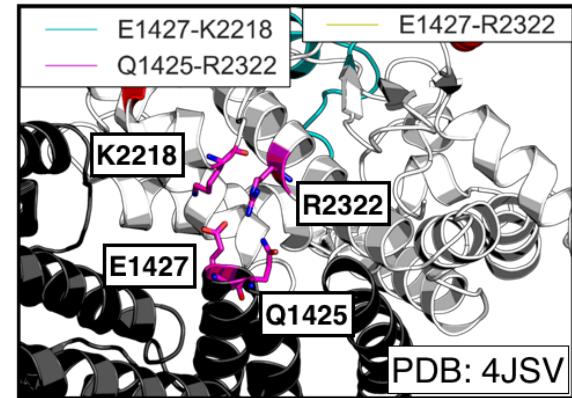
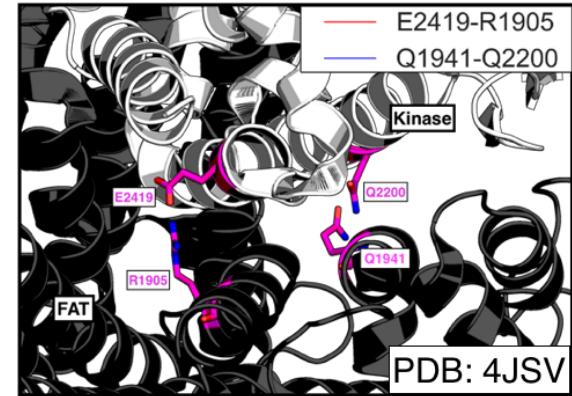
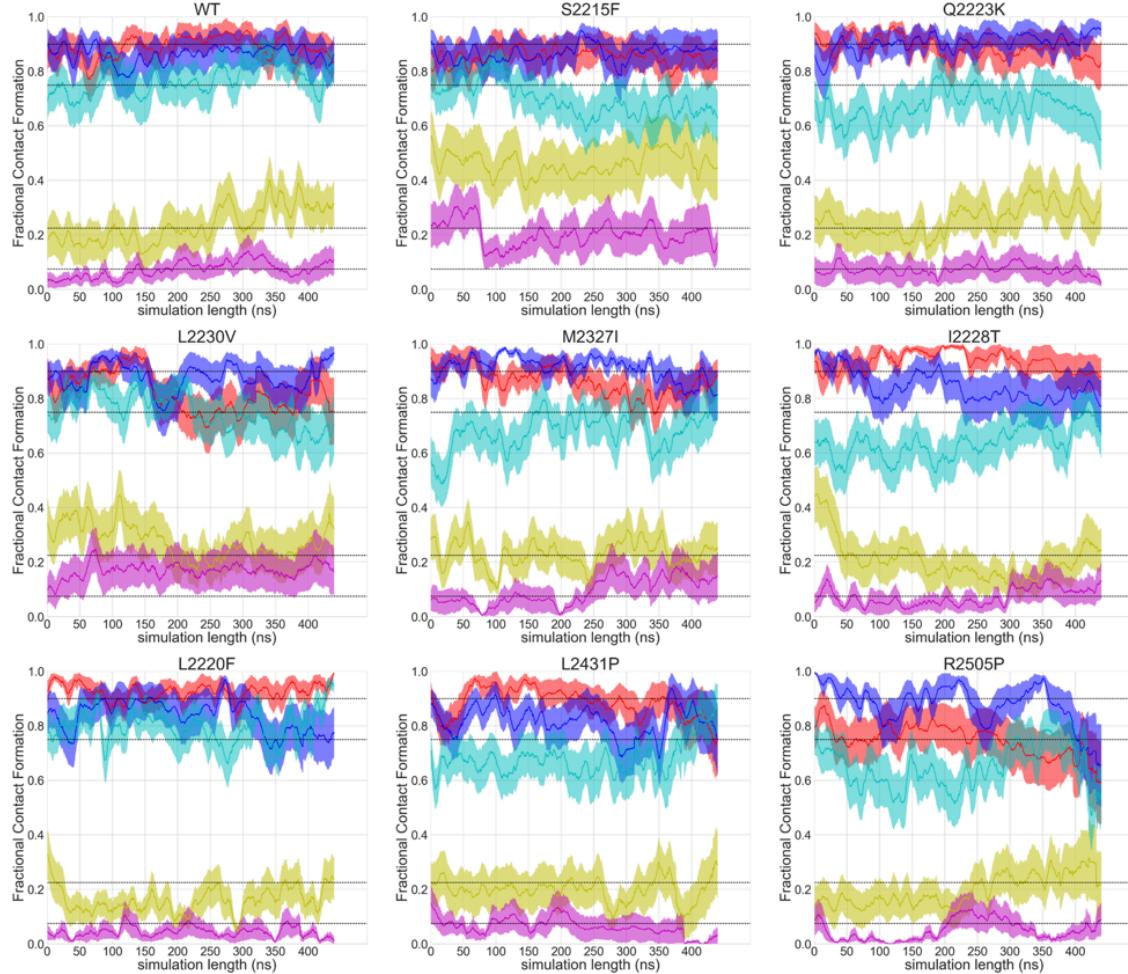


Figure 3.6: Missense mutations in kinase domain do not disrupt interactions between the kinase and FAT domains **Left** Fractional contact formation analysis for 10 500ns trajectories, analyzed in 20ns sliding window chunks. Contacts cutoff at 4 for closest heavy atom. Plotted is the mean \pm SEM. The dashed lines represent the proportion roughly populated in the WT simulations. **Right** Illustration of the distances being plotted between E2419 and R1905 (red), Q1941 and 2200 (blue), E1147 and K2218 (cyan), Q1425 and R2322 (magenta), or E1427 and R2322 (yellow). The kinase domain (white) interacts with the FAT domain (black) through salt bridges and hydrogen bonds formed by the highlighted residues (magenta)

3.3.4 Free energy calculations show promise in predicting impact of mutations on small molecule and ATP affinity

The preceding work on trying to understand the functional impact of mutations using traditional molecular dynamics highlights one of the key challenges of the application of physical modeling to this question: the need to compute meaningful, physically testable quantities. Alchemical free energy calculations have been previously used to predict the impact of mutations on protein-protein binding [175] and protein thermostabilities [176]. Based on this work, we set out to see if free energy calculations could be used to identify clinically observed mutations that might cause resistance to an ATP-competitive inhibitor. Here, we propose a model of resistance mutants that cause a decrease in affinity for an ATP-competitive inhibitor while maintaining a basic level of ATP affinity (Figure 3.7, left panels). An increase in ATP affinity could be a mechanism of resistance by making it more difficult for ATP-competitive inhibitors to out-compete ATP, as well as a mechanism of activation. To explore the impact of mutations on affinity for these compounds, we built models of mTOR bound to ATP (Figure 3.7, top left) and AZD8055 (Figure 3.7, bottom left). Using protein mutation FEP+ [103, 176], we ran preliminary calculations for a small panel of mutations from the MSKCC-IMPACT assay [177]. Three showed greater than 1-log unit, or 1.4 kcal/mol, decrease in affinity for AZD8055 (Figure 3.7, blue bars). W2239C, a missense mutation observed in a patient with cervical squamous cell carcinoma that occurred together with a known activating mutation S2215F [178, 179], showed the greatest decrease in affinity (shown as a positive $\Delta\Delta G_{mutation}$ in Figure 3.7). While we were not able to confirm via prediction that ATP-affinity is not maintained, this result suggests that treatment with an ATP-competitive inhibitor targeting mTOR may be less effective than expected.

M2345V and Y2225C are also both predicted to decrease the affinity for AZD8055. M2345V was observed in a cell line from a patient with cancer of unknown primary tissue and co-occurred with a TSC2 E75K mutation, which is likely to be loss of function [180]. Y2225C was observed in a patient with tubular stomach adenocarcinoma, who also had a slight copy number gain in both MTOR and DEPTOR [181]. All three of these mutations are exceedingly rare and occur in tumors with relatively high mutation burden. These preliminary results suggest that free energy calculations can potentially aid in understanding or teasing apart the functional impact of rare mutations in complex tumors.

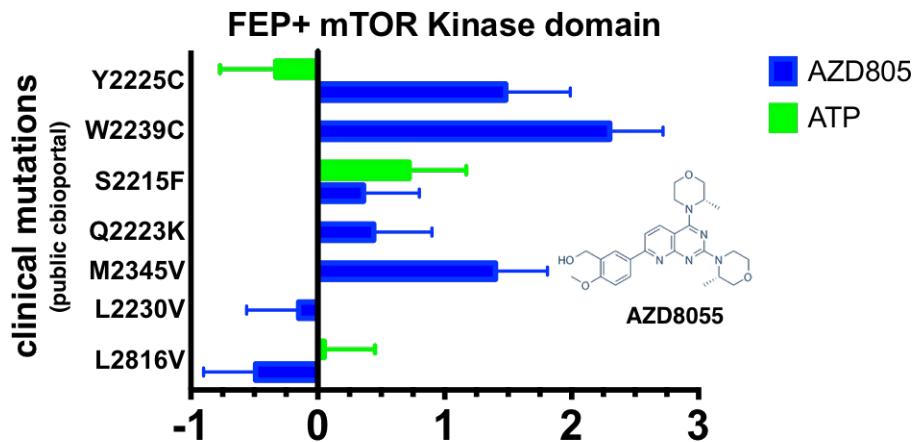
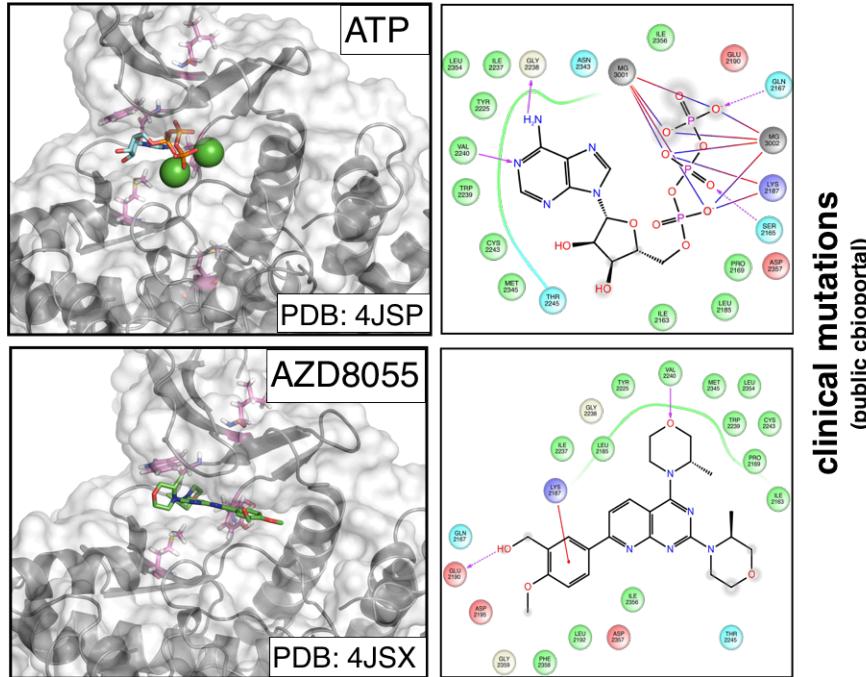


Figure 3.7: Free energy calculations identify potential resistance mutations **Left** Structures and 2D interactions maps of ATP (top, PDBID: 4JSP) and AZD8055 (bottom, PDBID: 4JSX) docked to the kinase domain of mTOR. Magnesium ions are shown as green spheres. **Right** $\Delta\Delta G_{mutation}$ (x-axis) calculated for a number of clinically observed mTOR mutations(y-axis) [177] for ATP (green) and AZD8055 (blue). Error bars correspond to the BAR uncertainty estimate.

3.4 Discussion and Conclusions

The above work utilizes physical modeling to study the impact of clinically observed mutations on mTOR structure, function and ligand binding. mTOR is a critical node in multiple signaling pathways with a complex structure and regulatory. Extensive molecular dynamics simulations were run on mutations that were confirmed to be hyperactivating by cellular and functional assays [152]. Contact map analysis suggests that local, fast structural rearrangements are induced by mutations such as S2215F. Unfortunately, more focused studies on order parameters of activation do not yield insight into potential mechanisms of activation, suggesting that either the simulation time is insufficient to observe changes in the activation state of the kinase domain or the mechanism of activation occurs at the level of mTORC or substrate access. These studies are also limited by the relative scarcity of biophysical data for mTOR. At the time of the study, there were no atomistic structures for the mTOR complexes, impeding any study of the role mTOR mutations play in disrupting or altering the structure of the complex. Recent, exciting work using Cryo-EM has yielded a number of high resolution structures for mTORC1 and mTORC2 [110, 182–184]. These advances would enable future work exploring the impact of mutations in the more physiologically relevant context of the complexes. Additionally, a more thorough understanding of the activation state of the mTOR crystal structure, which is assumed to be active, and what the inactive conformation of mTOR looks like would further enable these types of large scale studies.

The challenges with using traditional MD analysis highlights the importance of looking at testable, physical surrogates for activation and resistance. Unless there are well-validated order parameters from prior experimental work to use, drawing meaningful mechanistic conclusions from computation is difficult. Here,

we present a pilot study using alchemical free energy calculations to compute the impact of a small panel of clinical mutations on ATP and inhibitor affinity. Promisingly, we were able to identify three potential resistance mutations. Future work to introduce these mutations and measure binding affinities for AZD8055, as well as expanding the number of inhibitors and mutations in the dataset, could prove useful in understanding the functional impact of these mutations on treatment. Unfortunately, mTOR is a complicated and experimentally untractable system to work with, which motivated the studies presented in the next chapter of this work.

CHAPTER 4

PREDICTING THE IMPACT OF CLINICALLY-OBSERVED KINASE
MUTATIONS USING PHYSICAL MODELING

CHAPTER 5

ENABLING HIGH-THROUGHPUT BIOPHYSICAL EXPERIMENTS ON CLINICALLY-OBSERVED MUTATIONS

The work in this chapter was submitted to *Biochemistry* and appeared as:

An open library of human kinase domain constructs for automated bacterial expression

Steven K. Albanese^{†,1,2}, Daniel L. Parton^{†,2}, Mehtap Işık^{‡,2,3}, Lucelenie Rodríguez-Laureano^{‡,2}, Sonya M. Hanson², Julie M. Behr^{2,7}, Scott Gradia^{&,4}, Chris Jeans⁴, Nicholas M. Levinson⁶, Markus A. Seeliger⁵, John D. Chodera^{2,*}

¹ Louis V. Gerstner, Jr. Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, NY 10065

² Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065

³ Tri-Institutional PhD Program in Chemical Biology, Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, NY 10065

⁴ QB3 MacroLab, University of California, Berkeley, CA 94720

⁵ Department of Pharmacological Sciences, Stony Brook University Medical School, Stony Brook, NY 11794

⁶ Department of Pharmacology, University of Minnesota, Minneapolis, MN 55455

⁷ Tri-Institutional Program in Computation Biology and Medicine, Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, NY 10065

[†] or [‡] These authors contributed equally to this work

* Corresponding Author

& Present Address: Caribou Biosciences, Berkeley, CA 94720

5.1 Abstract

Kinases play a critical role in cellular signaling and are dysregulated in a number of diseases, such as cancer, diabetes, and neurodegeneration. Therapeutics targeting kinases currently account for roughly 50% of cancer drug discovery efforts. The ability to explore human kinase biochemistry and biophysics in the laboratory is essential to designing selective inhibitors and studying drug resistance. Bacterial expression systems are superior in terms of simplicity and

cost-effectiveness compared to insect or mammalian cells, but have historically struggled with human kinase expression. Following the discovery that phosphatase coexpression produced high yields of Src and Abl kinase domains in bacteria, we have generated a library of 52 His-tagged human kinase domain constructs that express above 2 μ g/mL culture in an automated bacterial expression system utilizing phosphatase coexpression (YopH for Tyr kinases, Lambda for Ser/Thr kinases). Here, we report a structural bioinformatics approach to identify kinase domain constructs previously expressed in bacteria and likely to express well in our protocol, experiments demonstrating our simple construct selection strategy selects constructs with good expression yields in a test of 84 potential kinase domain boundaries for Abl, and yields from a high-throughput expression screen of 96 human kinase constructs. Using a fluorescence-based thermostability assay and a fluorescent ATP-competitive inhibitor, we show that the highest-expressing kinases are folded and have well-formed ATP binding sites. We also demonstrate that these constructs can enable characterization of clinical mutations by expressing a panel of 48 Src and 46 Abl mutations. The wild-type kinase construct library is available publicly via Addgene.

5.2 Introduction

Kinases play a critical role in cellular signaling pathways, controlling a number of key biological processes that include growth and proliferation. There are over 500 kinases in the human genome [185, 186], many of which are of therapeutic interest. Perturbations due to mutation, translocation, or upregulation can cause one or more kinases to become dysregulated, often with disastrous

consequences [187]. Kinase dysregulation has been linked to a number of diseases, such as cancer, diabetes, and inflammation. Cancer alone is the second leading cause of death in the United States, accounting for nearly 25% of all deaths; in 2015, over 1.7 million new cases were diagnosed, with over 580,000 deaths [188]. Nearly 50% of cancer drug development is targeted at kinases, accounting for perhaps 30% of *all* drug development effort globally [83, 189].

The discovery of imatinib, an inhibitor that targets the Abelson tyrosine kinase (Abl) dysregulated in chronic myelogenous leukemia (CML) patients, was transformative in revealing the enormous therapeutic potential of selective kinase inhibitors, kindling hope that this remarkable success could be recapitulated for other cancers and diseases [190]. While there are now 39 FDA-approved selective kinase small molecule inhibitors (as of 16 Jan 2018) [3, 191], these molecules were approved for targeting only 22 out of ~500 human kinases¹, with the vast majority developed to target just a handful of kinases [192]. The discovery of therapeutically effective inhibitors for other kinases has proven remarkably challenging.

While these inhibitors have found success in the clinic, many patients cease to respond to treatment due to resistance caused by mutations in the targeted kinase [9], activation of downstream kinases [187], or relief of feedback inhibition in signaling pathways [193]. These challenges have spurred the development of a new generation of inhibitors aimed at overcoming resistance [34, 35], as well as mutant-specific inhibitors that target kinases bearing a missense mutation that confers resistance to an earlier generation inhibitor [36]. The ability to easily engineer and express mutant kinase domains of interest would be of enormous aid in the development of mutant-selective inhibitors, offering an advantage over

¹These targets are, currently: Abl, DDR1, EGFR, HER2, VGFR1/2/3, Alk, Met, BRAF, JAK1/2/3, Btk, Pi3K, CDK4, CDK6, MEK, ROS1, FLT3, IGF1R, Ret, Kit, Axl, TrkB, and mTOR [3].

current high-throughput assays [194–196], which typically include few clinically-observed mutant kinases.

Probing human kinase biochemistry, biophysics, and structural biology in the laboratory is essential to making rapid progress in understanding kinase regulation, developing selective inhibitors, and studying the biophysical driving forces underlying mutational mechanisms of drug resistance. While human kinase expression in baculovirus-infected insect cells can achieve high success rates [197, 198], it cannot compete in cost, convenience, or speed with bacterial expression. *E. coli* expression enables production of kinases without unwanted post-translational modifications, allowing for greater control of the system. A survey of 62 full-length non-receptor human kinases found that over 50% express well in *E. coli* [197], but often expressing only the soluble kinase domains are sufficient, since these are the molecular targets of therapy for targeted kinase inhibitors and could be studied even for receptor-type kinases. While removal of regulatory domains can negatively impact expression and solubility, coexpression with phosphatase was shown to greatly enhance bacterial kinase expression in Src and Abl tyrosine kinases, presumably by ensuring that kinases remain in an unphosphorylated inactive form where they can cause minimal damage to cellular machinery [199].

The protein databank (PDB) now contains over 100 human kinases that were expressed in bacteria, according to PDB header data. Many of these kinases were expressed and crystallized as part of the highly successful Structural Genomics Consortium (SGC) effort to increase structural coverage of the human kinome [200]. Since bacterial expression is often complicated by the need to tailor construct boundaries, solubility-promoting tags, and expression and purification protocols individually for each protein expressed, we wondered whether a

simple, uniform, automatable expression and purification protocol could be used to identify tractable kinases, select construct boundaries, express a large number of human kinases and their mutant forms, and produce a convenient bacterial expression library to facilitate kinase research and selective inhibitor development. As a first step toward this goal, we developed a structural informatics pipeline to use available kinase structural data and associated metadata to select constructs from available human kinase libraries to clone into a standard set of vectors intended for phosphatase coexpression under a simple automatable expression and purification protocol. Using an expression screen for multiple construct domain boundaries of Abl, we found that transferring construct boundaries from available structural data can produce constructs with useful expression levels, enabling simple identification of construct domain boundaries. We then completed an automated expression screen in Rosetta2 cells of 96 different kinases and found that 52 human kinase domains express with yields greater than 2 µg/mL culture. To investigate whether these kinases are properly folded and useful for biophysical experiments, we performed a fluorescence-based thermostability assay on the 14 highest expressing kinases in our panel and a single-well high-throughput fluorescence-based binding affinity measurement on 39 kinases. These experiments demonstrated that many of the expressed kinases were folded, with well formed ATP binding sites capable of binding a small molecule kinase inhibitor. To demonstrate the utility of these constructs for probing the effect of clinical mutations on kinase structure and ligand binding, we subsequently screened 48 Src and 46 Abl mutations, finding that many clinically-derived mutant kinase domains can be expressed with useful yields in this uniform automated expression and purification protocol.

All source code, data, and wild-type kinase plasmids associated with this project are freely available online:

- **Source code and data:**
<https://github.com/choderalab/kinase-ecoli-expression-panel>
- **Interactive table of expression data:**
<http://choderalab.org/kinome-expression>
- **Plasmids:** <https://www.addgene.org/kits/chodera-kinase-domains>

5.3 Results

5.3.1 Construct boundary choice impacts Abl kinase domain expression

To understand how alternative choices of expression construct boundaries can modulate bacterial expression of a human kinase domain, we carried out an expression screen of 84 unique construct boundaries encompassing the kinase domain of the tyrosine protein kinase ABL1.

Three constructs known to express in bacteria were chosen from the literature and used as controls, spanning Uniprot residues 229–500 (PDBID: 3CS9) [201], 229–512 (PDBID: 2G2H) [202] and 229–515 (PDBID: 2E2B) [203]. 81 constructs were generated combinatorially by selecting nine different N-terminal boundaries spanning residues 228–243 and nine different C-terminal boundaries spanning residues 490–515, chosen to be near the start and end points for the control constructs (Figure 5.1A). Each of the three control constructs included six replicates to provide an estimate of the typical standard error in expression readout for the experimental constructs, which was found to be between 0.42–1.5 µg/mL (Figure 5.1A, green constructs).

Briefly, the impact of construct boundary choice on Abl kinase domain expression was tested as follows (see Methods for full details). His10-TEV N-terminally tagged wild-type Abl constructs² were coexpressed with YopH phosphatase in a 96-well format with control replicates distributed randomly throughout the plate. His-tagged protein constructs were recovered via a single nickel affinity chromatography step, and construct yields were quantified using microfluidic capillary electrophoresis following thermal denaturation. Expression yields are summarized in Figure 5.1A, and a synthetic gel image from the constructs with detectable expression is shown in Figure 5.2. Abl construct bands are present at sizes between 29 and 35 kDa (due to the variation in construct boundaries), and YopH phosphatase (which is not His-tagged but has substantial affinity for the nickel beads) is present in all samples at its expected size of 50 kDa. Strikingly, despite the fact that N-terminal and C-terminal construct boundaries only varied over 15–25 residues, only a small number of constructs produced detectable expression (Figure 5.1B). As highlighted in Figure 5.1C (left), the best N-terminal boundaries (residues 228, 229, 230) are located on a disordered strand distant from any secondary structure; N-terminal boundaries closer to the beta sheet of the N-lobe gave poor or no detectable expression (Figure 5.1B).

The best C-terminal construct boundaries (residues 511 and 512) occur in an α -helix (Figure 5.1C, right). Of note, this α -helix is not resolved in PDBID:2E2B [203], suggesting this structural element may only be weakly thermodynamically stable in the absence of additional domains. In previous work, this α -helix was shown to undergo a dramatic conformational change which introduces a kink at residue 516, splitting the α -helix into two [20]. This suggests a high potential for flexibility in this region.

²Parent plasmid is a pET His10 TEV LIC cloning vector and is available on Addgene (Plasmid #78173).

Two of the control constructs (which differ in construct boundary by only one or two residues) were in the top six expressing constructs (Figure 5.1A), and were in fact within 60% of the maximum observed expression yield. From this, we concluded that transferring construct boundaries from existing kinase domain structural data would be sufficient to bias our constructs toward useful expression levels for a large-scale screen of multiple kinases.

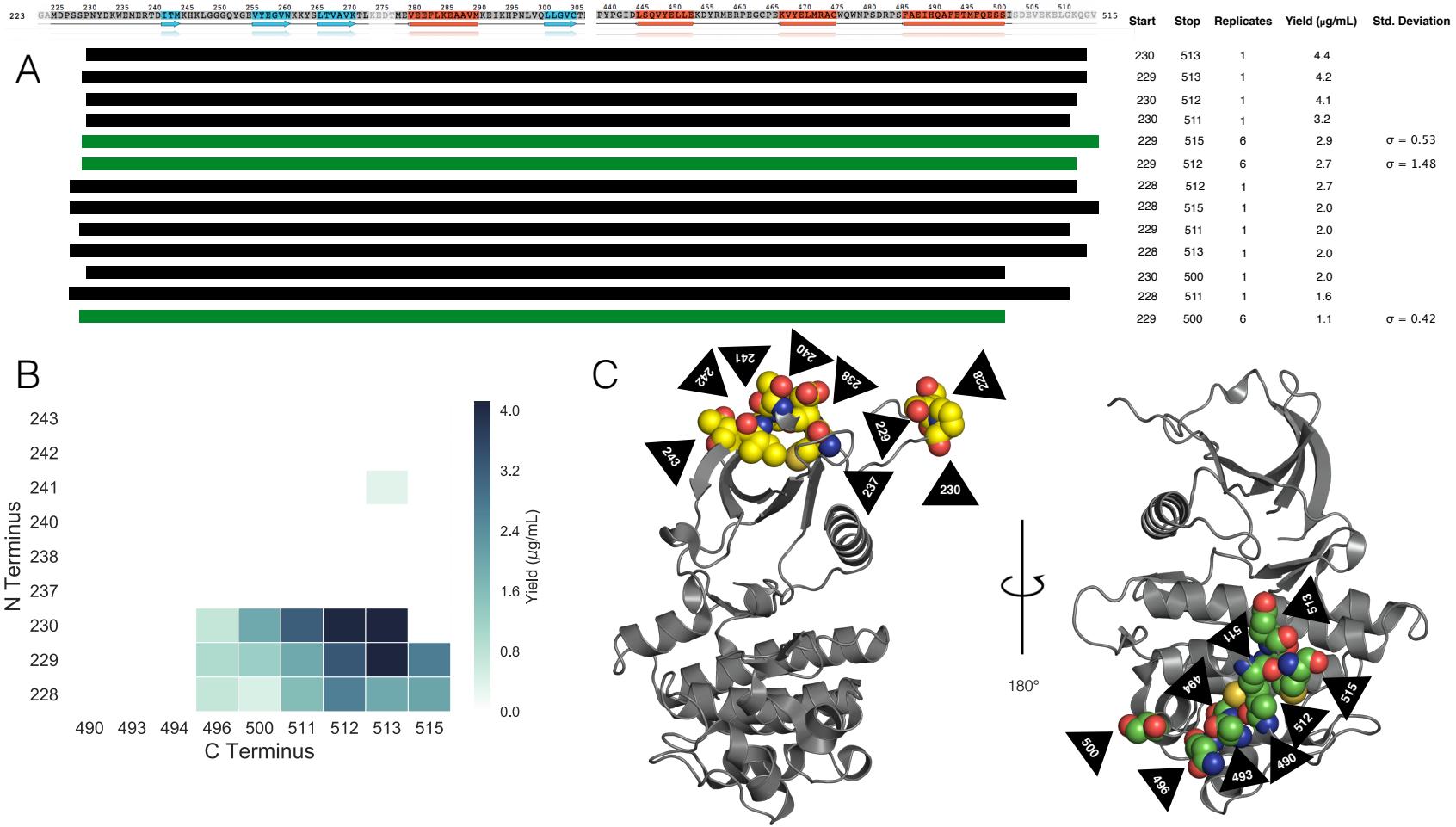


Figure 5.1: Abl kinase domain construct expression screen illustrates high sensitivity to construct boundaries. (A) Abl kinase domain construct boundaries with highest expression yields. Standard deviations of the yield are listed for control constructs for which six replicates were performed to give an indication of the uncertainty in experimental constructs. Secondary structure is indicated on the sequence. Beta sheets are colored blue and alpha helices are colored orange. (B) Heatmap showing average yields for constructs (in $\mu\text{g/mL}$ culture) with detectable expression as a function of N- and C-terminal construct boundaries. (C) left: PDBID: 2E2B with the nine N-terminal construct boundary amino acids shown as yellow spheres. right: PDBID: 4XEY with the nine C-terminal construct boundary amino acids shown as green spheres. Black arrows indicate residue numbers.

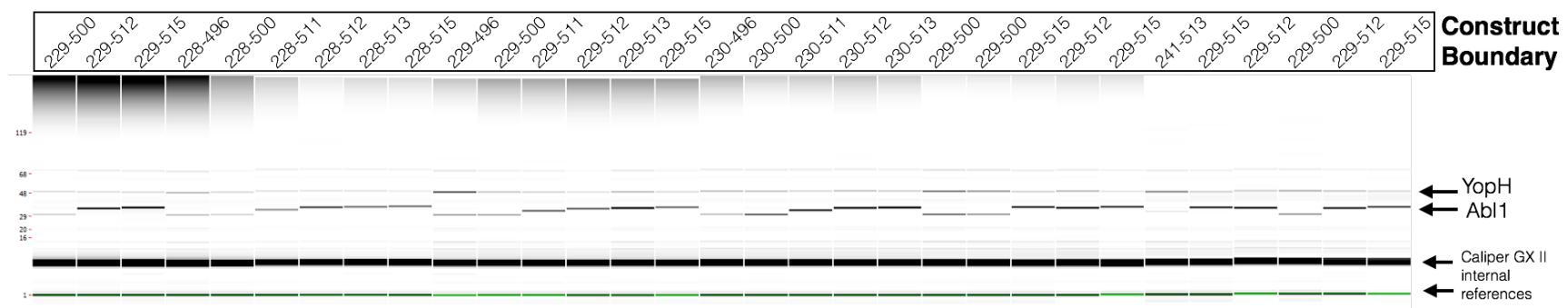


Figure 5.2: **Expression yields of Abl kinase domain constructs for all constructs with detectable expression.** A synthetic gel image rendering generated from Caliper GX II microfluidic gel electrophoresis data following Ni-affinity purification and thermal denaturation for all Abl constructs with detectable expression. Each well is marked with the Abl kinase domain construct residue boundaries (Uniprot canonical isoform numbering). Bands for YopH164 phosphatase (50 kDa) and Abl kinsase domain constructs (28–35 kDa) are labeled.

5.3.2 Screen of 96 kinases finds 52 with useful levels of automated *E. coli* expression

To begin exploring which human kinase domains can achieve useful expression in *E. coli* using a simple automatable expression and purification protocol, a panel of kinase domain constructs for 96 kinases, for which bacterial expression has been previously demonstrated, was assembled using a semi-automated bioinformatics pipeline. Briefly, a database was built by querying Uniprot [204] for human protein kinase domains that were both active and not truncated. This query returned a set of target sequences that were then matched to their relevant PDB constructs and filtered for expression system (as determined from PDB header EXPRESSION_SYSTEM records), discarding kinases that did not have any PDB entries with bacterial expression. As a final filtering step, the kinases were compared to three purchased kinase plasmid libraries (described in Methods), discarding kinases without a match. Construct boundaries were selected from PDB constructs and the SGC plasmid library, both of which have experimental evidence for *E. coli* expression, and subcloned from a plasmid in a purchased library (see Methods). Selecting the kinases and their constructs for this expression trial in this method rested on the basis of expected success: these specific kinase constructs were bacterially expressed and purified to a degree that a crystal structure could be solved. While expression protocols used to produce protein for crystallographic studies are often individually tailored, we considered these kinases to have a high likelihood of expressing in our semi-automated pipeline where the *same* protocol is utilized for all kinases. Statistics of the number of kinases obtained from the PDB mining procedure are shown in Figure 5.3A. Surprisingly, the most highly sampled family was the CAMK family, suggesting researchers may have found this family particularly amenable to

bacterial expression. Based on the results of the previous experiment scanning Abl constructs for expression, we decided to use construct boundaries that were reported in the literature for each kinase. This process resulted in a set of 96 plasmid constructs distributed across kinase families (Figure 5.3B).

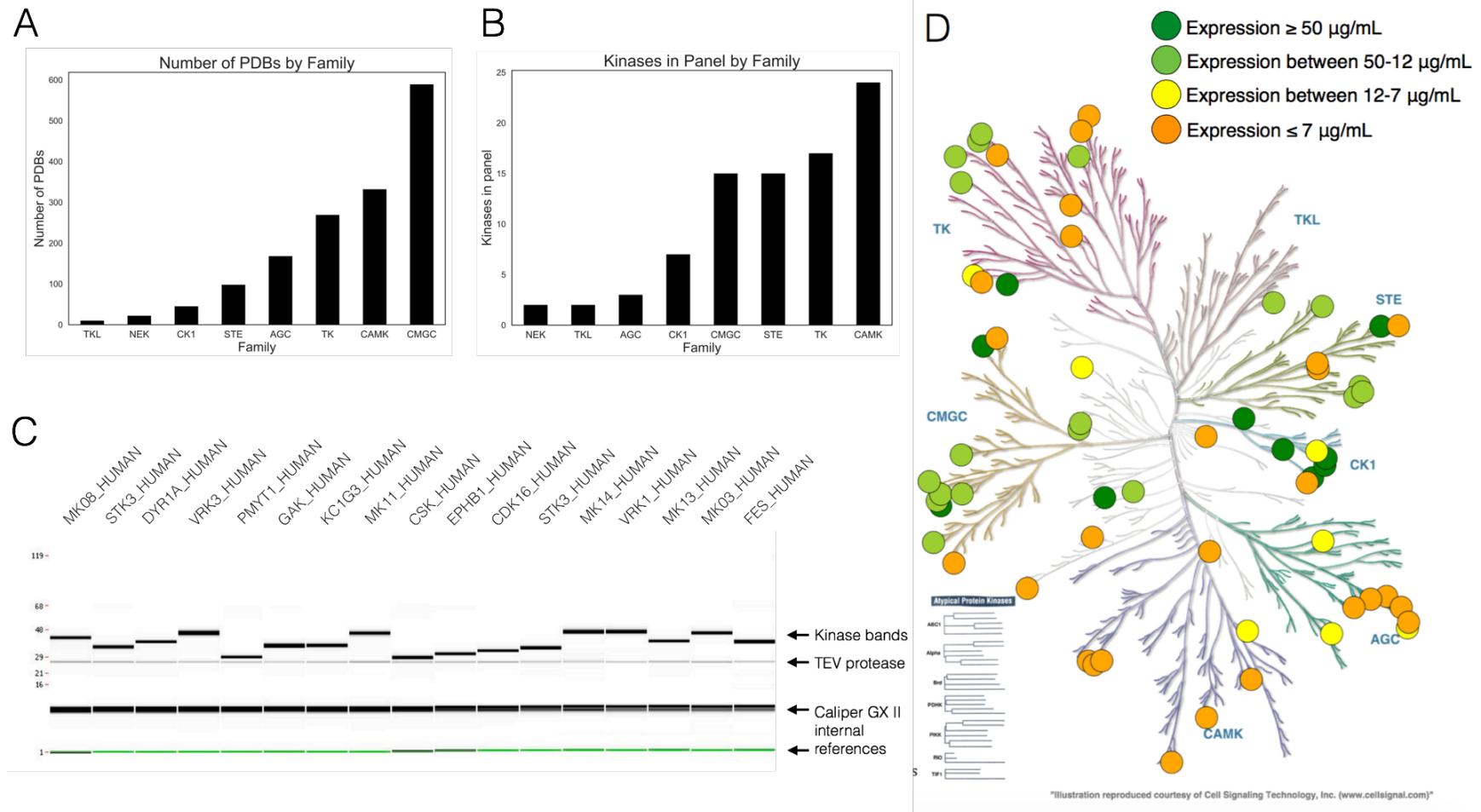


Figure 5.3: Kinome wide search for expressible kinases. (A) The number of PDB structures per kinase family, from the database built to select kinases for expression. (B) The distribution among families of candidate kinases in our expression screen. (C) Caliper GX II synthetic gel image rendering of the highest expressing kinases, quantified using microfluidic capillary electrophoresis. (D) Kinome distribution of expression based on our 96 kinase screen. Dark green circles represent kinases with expression above 50 $\mu\text{g}/\text{mL}$ culture yield. Light green circles represent kinases with expression between 50 and 12 $\mu\text{g}/\text{mL}$ yield. Yellow circles represent kinases with expression between 12 and 7 $\mu\text{g}/\text{mL}$ yield. Orange circles represent kinases with any expression (even below 2 $\mu\text{g}/\text{mL}$) up to 7 $\mu\text{g}/\text{mL}$ yield. Image made with KinMap: <http://www.kinhub.org/kinmap>.

Table 5.1: Kinase domain constructs with yields $\geq 2 \mu\text{g/mL}$ culture for 96-kinase expression screen. Kinases are listed by Uniprot designation and whether they were co-expressed with Lambda or truncated YopH164 phosphatase. Yield (determined by Caliper GX II quantitation of the expected size band) reported in $\mu\text{g/mL}$ culture, where total eluate volume was $120 \mu\text{L}$ from $900 \mu\text{L}$ bacterial culture. Yields are shaded green (yield $\geq 12 \mu\text{g/mL}$), yellow ($12 > \text{yield} \geq 7 \mu\text{g/mL}$) and orange (yield $< 7 \mu\text{g/mL}$); kinase domain constructs with yields that were undetectable or $< 2 \mu\text{g/mL}$ are not listed. \ddagger denotes that the second kinase domain of KS6A1_HUMAN was expressed; all other kinases were the first or only kinase domain occurring in the ORF. Construct boundaries are listed in UniProt residue numbering for the UniProt canonical isoform. An interactive table of expression yields and corresponding constructs is available at <http://choderlab.org/kinome-expression>

Kinase	Construct Boundary	Plasmid Source and ID	Phosphatase	Yield ($\mu\text{g/mL}$)
MK14_HUMAN	1–360	Addgene 23865	Lambda	70.7
VRK3_HUMAN	24–352	SGC Oxford VRK3A-c016	Lambda	67.5
GAK_HUMAN	24–359	SGC Oxford GAKA-c006	Lambda	64.7
CSK_HUMAN	186–450	Addgene 23941	YopH	62.5
VRK1_HUMAN	3–364	Addgene 23496	Lambda	62.3
KC1G3_HUMAN	24–351	SGC Oxford CSNK1G3A-c002	Lambda	56.3
FES_HUMAN	448–822	Addgene 23876	YopH	44.0
PMYT1_HUMAN	24–311	SGC Oxford PKMYT1A-c004	Lambda	38.0
MK03_HUMAN	1–379	Addgene 23509	Lambda	36.4
STK3_HUMAN	16–313	Addgene 23818	Lambda	34.3
DYR1A_HUMAN	24–382	SGC Oxford DYRK1AA-c004	Lambda	34.1
KC1G1_HUMAN	24–331	SGC Oxford CSNK1G1A-c013	Lambda	34.1
MK11_HUMAN	24–369	SGC Oxford MAPK11A-c007	Lambda	31.7
MK13_HUMAN	1–352	Addgene 23739	Lambda	31.7
EPHB1_HUMAN	602–896	Addgene 23930	YopH	28.9
MK08_HUMAN	1–363	HIP pJP1520 HsCD00038084	Lambda	28.5
CDK16_HUMAN	163–478	Addgene 23754	Lambda	26.9
EPHB2_HUMAN	604–898	HIP pJP1520 HsCD00038588	YopH	25.1
PAK4_HUMAN	291–591	Addgene 23713	Lambda	23.9
CDKL1_HUMAN	2–304	SGC Oxford CDKL1A-c024	Lambda	23.2
SRC_HUMAN	254–536	Addgene 23934	YopH	22.0
STK16_HUMAN	24–316	SGC Oxford STK16A-c002	Lambda	20.7
MAPK3_HUMAN	33–349	Addgene 23790	Lambda	18.8
PAK6_HUMAN	383–681	Addgene 23833	Lambda	18.0

Table 5.1 continued from previous page

Kinase	Construct Boundary	Plasmid Source and ID	Phosphatase	Yield (μ g/mL)
CSK22_HUMAN	1–334	HIP pJP1520 HsCD00037966	Lambda	17.9
MERTK_HUMAN	570–864	Addgene 23900	YopH	16.8
PAK7_HUMAN	24–318	SGC Oxford PAK5A-c011	Lambda	14.7
CSK21_HUMAN	1–335	Addgene 23678	Lambda	14.5
EPHA3_HUMAN	606–947	Addgene 23911	YopH	14.1
BMPR2_HUMAN	1–329	SGC Oxford BMPR2A-c019	Lambda	14.1
M3K5_HUMAN	659–951	HIP pJP1520 HsCD00038752	Lambda	14.0
KCC2G_HUMAN	24–334	SGC Oxford CAMK2GA-c006	Lambda	13.3
E2AK2_HUMAN	254–551	HIP pJP1520 HsCD00038350	Lambda	11.6
MK01_HUMAN	1–360	HIP pJP1520 HsCD00038281	Lambda	11.2
CSKP_HUMAN	1–340	HIP pJP1520 HsCD00038384	Lambda	10.1
CHK2_HUMAN	210–531	Addgene 23843	Lambda	8.1
KC1G2_HUMAN	4–312	SGC Oxford CSNK1G2A-c002	Lambda	7.6
DMPK_HUMAN	24–433	SGC Oxford DMPK1A-c026	Lambda	7.6
KCC2B_HUMAN	11–303	Addgene 23820	Lambda	7.1
FGFR1_HUMAN	456–763	Addgene 23922	YopH	6.1
KS6A1_HUMAN [‡]	413–735	SGC Oxford RPS6KA1A-c036	Lambda	5.7
DAPK3_HUMAN	9–289	Addgene 23436	Lambda	4.0
STK10_HUMAN	18–317	HIP pJP1520 HsCD00038077	Lambda	3.7
KC1D_HUMAN	1–294	Addgene 23796	Lambda	3.7
KC1E_HUMAN	1–294	Addgene 23797	Lambda	3.5
NEK1_HUMAN	23–350	SGC Oxford NEK1A-c011	Lambda	3.3
CDK2_HUMAN	1–297	Addgene 23777	Lambda	3.1
ABL1_HUMAN	229–512	HIP pJP1520 HsCD00038619	YopH	2.5
DAPK1_HUMAN	2–285	HIP pJP1520 HsCD00038376	Lambda	2.4
DYRK2_HUMAN	23–417	SGC Oxford DYRK2A-c023	Lambda	2.4
HASP_HUMAN	24–357	SGC Oxford GSG2A-c009	Lambda	2.3
FGFR3_HUMAN	449–759	Addgene 23933	YopH	2.3

From these constructs, a set of 96 His10-TEV N-terminally tagged kinase domain constructs were generated, coexpressed with a phosphatase in *E. coli*, purified via nickel bead pulldown, and quantified using microfluidic gel electrophoresis. The 96 kinases were coexpressed with either Lambda phosphatase (for Ser/Thr kinases) or a truncated form of YopH phosphatase³ (for Tyr kinases).

Instead of eluting with imidazole, purified kinase was cleaved off nickel beads by the addition of 10% TEV protease to minimize phosphatase contamination in the resulting eluate, allowing us to assess whether resulting yields would be sufficient (and sufficiently free of phosphatase) to permit activity assays. While the initial panel of 96 kinases was well-distributed among kinase families (Figure 5.3B), the most highly expressing kinases (yield of more than 12 μg kinase/mL culture) were not evenly distributed (Figure 5.3D). While many of the kinases chosen from the CMGC and CK1 families expressed well in our panel, nearly all of the kinases from the CAMK and AGC family express below 12 μg kinase/mL (Figure 5.3D). 52 kinases demonstrated a useful level of soluble protein expression, here defined as greater than 2 $\mu\text{g}/\text{mL}$, naïvely expected to scale up to better than 2 mg/L culture (Table 5.1). Some kinases (shaded green in Table 5.1) demonstrated very high levels of expression, while others (shaded orange in Table 5.1) would likely benefit from further rounds of construct boundary optimization or solubility tags to boost soluble expression. The 17 most highly expressing kinases showed relatively high purity after elution, though we note that eluting via TEV site cleavage results in a quantity of TEV protease in the eluate (Figure 5.3C), but does not cause the elution of the His-tagged phosphatases which would hinder the ability to perform kinase activity assays. Further optimization of elution conditions may be required for optimizing kinase

³Yoph164 phosphatase, engineered to minimize intrinsic affinity for nickel purification resin by the QB3 MacroLab based on parent plasmid pCDFDuet1-YOPH, a gift from the Kuriyan Lab.

recovery via TEV cleavage [205–207].

Constructs with expression yields above 2 µg/mL have been made available via
Addgene: <https://www.addgene.org/kits/chodera-kinase-domains>

5.3.3 High-expressing kinases are folded with a well-formed ATP binding site

To determine whether the expressed kinases were properly folded, we performed both a fluorescence-based thermostability assay (Figure 5.4) as well as a fluorescent ATP-competitive ligand binding measurement to quantify whether the ATP binding site was well-formed (Figure 5.5).

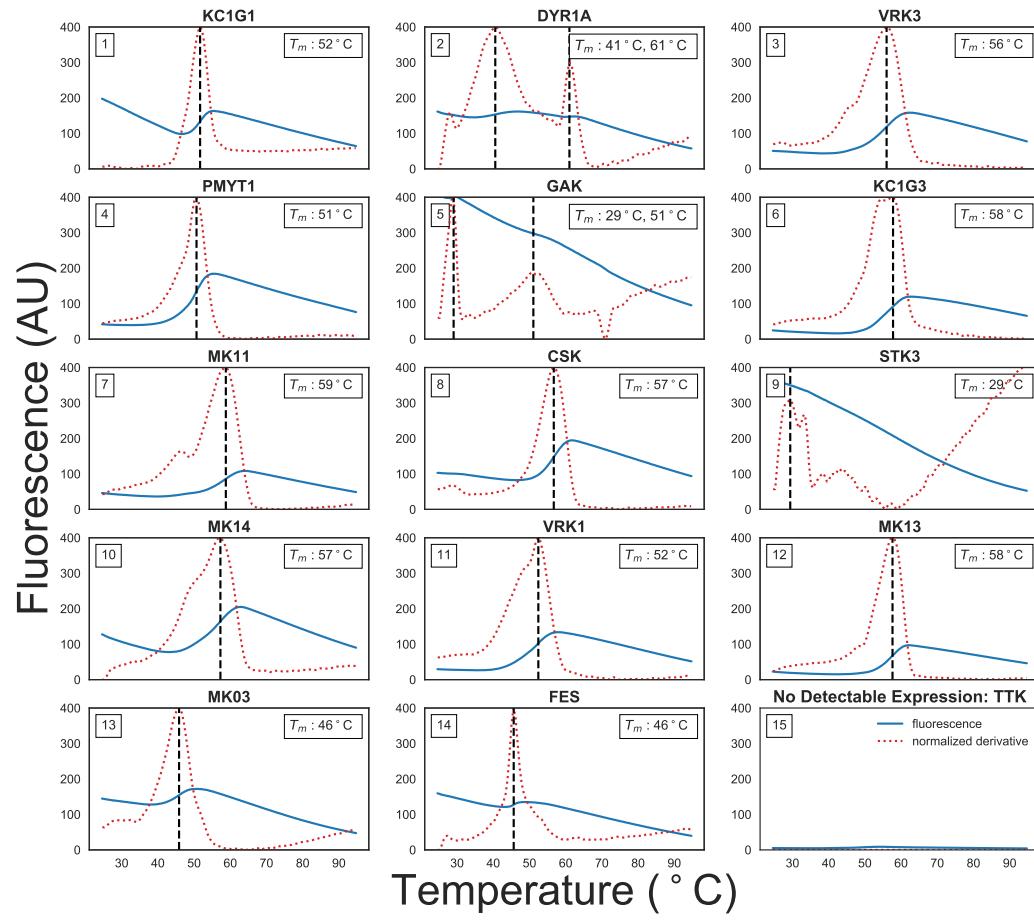


Figure 5.4: Fluorescence-based thermostability assay demonstrates many high-expressing kinases are well-folded. A fluorescence-based thermostability assay was performed on the 14 kinases shown to express above a minimum 0.24 mg/mL concentration after elution. SYPRO Orange fluorescence (solid blue line) was measured at 580 nm (half bandwidth 20 nm) after excitation at 465 nm (half bandwidth 25 nm) as the temperature was ramped from (x-axis) in Nickel Buffer A (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl, 20 mM imidazole, 1 mM BME). The temperature was held at 25°C for 15 sec before ramping up to 95°C with a ramp rate of 0.06°C/s. The unfolding temperature T_m (black dashed line and insert) was determined from the maxima of the normalized first derivative of fluorescence (red dashed line). Fluorescence emission at 580 nm is shown on the left y-axis. To control for signals resulting from TEV protease contamination present at 0.01–0.03 mg/mL, TTK, a kinase with no detectable expression in our panel as determined via Caliper GX II quantitation was included (panel 15).

5.3.4 Fluorescence-based thermostability assay

A fluorescence-based thermostability assay was performed with the hydrophobic dye SYPRO Orange to determine whether a strong two-state unfolding signal could be observed (see Methods). Also referred to as *thermofluor* or *differential scanning fluorimetry (DSF)*, as the temperature is slowly increased, unfolded proteins will expose hydrophobic patches that SYRPO orange will bind to, causing an increase in fluorescence [208–210]. While the fluorescence of solvated SYPRO Orange is temperature-dependent, clear unfolding temperatures (T_m) can often be identified from peaks in the first derivative of the observed fluorescence signal. Figure 5.4 shows the fluorescence (blue line), the absolute value of its derivative (red dashed line), and the unfolding temperature determined from the maximum absolute derivative (T_m) for the the 14 kinases that were eluted to concentrations above 0.24 mg/mL eluate, which was determined to be the minimum concentration required for optimal resolution of melting curves upon dilution to 10 μ L. Because TEV-eluted kinase was used directly in this assay, TEV protease contaminant varies from 0.01–0.03 mg/mL in the resulting assay mix. The selected minimum concentration ensured that the kinase was roughly an order of magnitude higher concentration than the contaminating TEV.

Most of the kinases assayed had strong peaks above room temperature, suggesting that they are well-folded in the elution buffer (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl, 20 mM imidazole, 1 mM BME) at room temperature. Some kinases, such as a DYR1A and GAK (Figure 5.4, panels 6 and 9), had two shallow inflection points in SYPRO fluorescence as a function of temperature. While STK3 does not have a strong peak above room temperature, titration with an ATP-competitive inhibitor suggests this kinase either has a well-formed ATP binding site or folding can be induced by ligand binding (Figure 5.5, panel 10). As

a control, a sample with no detectable kinase expression (TTK from our expression panel) was assayed (Figure 5.4, panel 9), which showed nearly no fluorescence signal.

5.3.5 ATP-competitive inhibitor binding fluorescence assay

To determine whether expressed kinases had well-folded ATP binding sites, we probed their ability to bind an ATP-competitive inhibitor. While a pan-kinase inhibitor such as staurosporine could be used as a fluorescent probe [211], the ATP-competitive inhibitor bosutinib shows a much stronger increase in fluorescence around 450–480 nm when bound to kinases with well-folded ATP binding sites [212, 213]. While excitation at 350 nm can be used, excitation at 280 nm results in lower background, potentially due to fluorescent energy transfer between kinase and ligand. Despite the weak affinity of bosutinib for many kinases, its aqueous solubility is sufficient to provide a quantitative assessment of ATP-competitive binding to many kinases at sufficiently high concentrations to function as a useful probe [212, 213].

Here, we utilized this approach as a *qualitative* probe for ATP-competitive ligand binding, due to uncertainty in the ligand concentration caused by significant evaporation over the course of the sequential titration experiment (see Methods section for a more in depth discussion). 33 of the kinases in our expression panel had sufficient yields to prepare 100 μ L of 0.5 μ M kinase assay solutions, and were assessed for binding to bosutinib (Figure 5.5, panels 1-33), with a concentration-dependent increase in fluorescence signal (colored spectra) over the baseline ligand fluorescence titrated into buffer (gray spectra) providing evidence of a well-formed ATP binding site. Six of the lowest expression kinase constructs

(Figure 5.5, panels 39-44) were prepared by diluted 20 μ L to a reaction volume of 100 μ L and assessed for bosutinib binding. Unexpectedly, these kinases also showed evidence of binding, suggesting this assay is able to detect a well-formed ATP binding site even for protein concentrations less than 0.5 μ M. To demonstrate that unfolded kinases do not demonstrate this increase in fluorescence over ligand-only baseline, thermally denatured MK14 was included as a control next to folded MK14 from a large-scale expression prep (Figure 5.5, panels 37–38), with thermally denatured MK14 exhibiting little difference from titrating ligand into buffer alone.

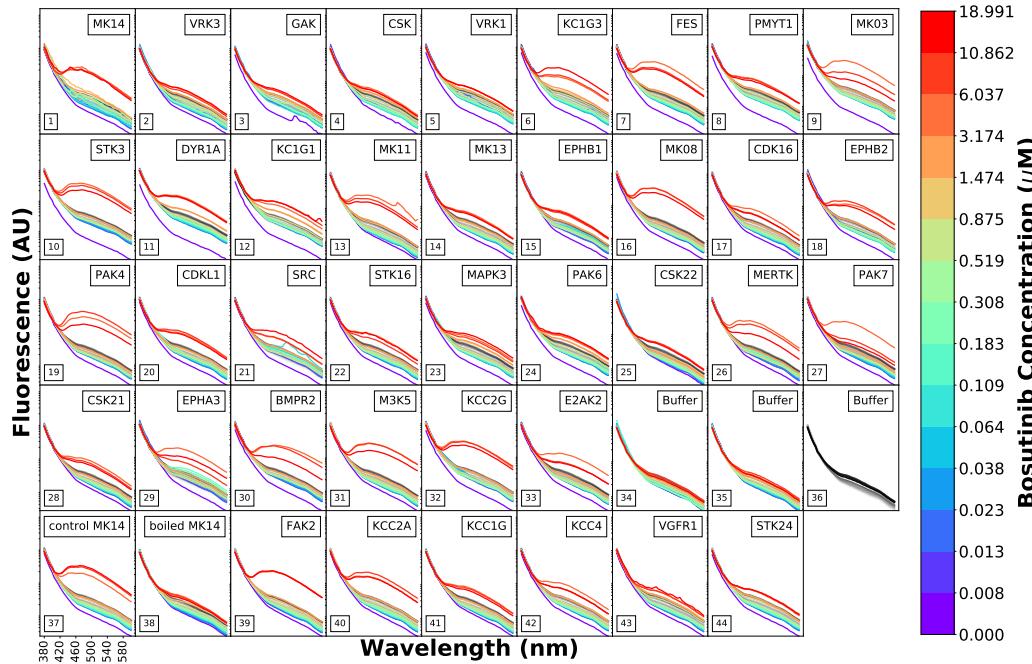


Figure 5.5: Fluorescence emission spectra as a function of the fluorescent ATP-competitive kinase inhibitor bosutinib demonstrates the presence of a well-formed ATP binding pocket. The ATP-competitive inhibitor bosutinib shows a strong increase in fluorescence centered around 450 nm when bound to kinases with well-folded ATP binding sites upon excitation at 280 nm [212]. To assess whether the kinases from the high-throughput expression screen were well-folded, bosutinib was titrated in a 15-concentration series geometrically spanning 0.008 μM to 18.99 μM (colored lines, higher concentrations are shown in warmer colors) in 15 increments for 39 expressing kinases with protein concentration adjusted to $\sim 0.5 \mu\text{M}$ in 100 μL assay volume. Eluted TEV protease contaminant varies from 0.01–0.03 mg/mL in the assay volumes. The control MK14 and boiled MK14 (boiled for 10 min at 95°C) were produced in a large scale expression from the same plasmid as used in the high-throughput expression protocol and they were included as positive and negative controls for bosutinib binding to ATP binding pocket. Fluorescence emission spectra (y-axis, bandwidth 20 nm) were measured from 370 nm to 600 nm (x-axis) for excitation at 280 nm (bandwidth 10 nm). For reference, the fluorescence of bosutinib titrated into buffer titration (panel 36) is shown in grayscale in each panel. Significant increases in fluorescence signal above baseline qualitatively indicate the presence of a well-formed ATP binding site.

5.3.6 Expressing clinically-derived Src and Abl mutants

Next-generation sequencing has enabled generation of massive datasets rich with missense alterations in kinases observed directly in the clinic [37–39], and has been particularly transformative in the field of oncology. To determine how well our human kinase domain panel supports the automated expression of clinically-identified missense mutants for biophysical, biochemical, and structural characterization, we attempted to express 96 missense mutations mined from sequencing studies of cancer patients. The mutations were gathered using cBioPortal [214] from publicly available sources and a large clinical tumor sequencing dataset from the Memorial Sloan Kettering Cancer Center [38] sequenced in the MSK-IMPACT panel [215].

Table 5.2: Expression yields for engineered clinical missense mutants of Abl kinase domains with yields $\geq 2 \mu\text{g/mL}$ culture. Abl kinase domain constructs with engineered clinical mutations with expression yields $\geq 2 \mu\text{g/mL}$ culture are listed, sorted by yield. Yield was determined by Caliper GX II quantitation of the expected size band and reported in $\mu\text{g/mL}$ culture, where total eluate volume was 80 μL purified from 900 μL bacterial culture. Wild-type (WT) controls for both Src and Abl (here, a single well for each) are shown as the first entry for each gene.

Abl1 (229–512)	Mutation ^a	Functional Impact Score ^b	yield ($\mu\text{g/mL}$)	% of WT expression
WT	–	5.1	–	–
I403T	Low	17.8	350	
I293M	Low	9.8	193	
P309S	Neutral	7.8	153	
E453K	Low	7.3	144	
Y440H	Medium	7.1	140	
E292D	Low	6.9	135	
G251C	High	5.2	102	
E282Q	Neutral	5.1	102	
G250R	Neutral	5.1	100	
G254R	High	5.0	98	
Y312C	Neutral	4.7	93	
E453Q	Low	3.7	73	
R328K	Low	3.5	69	
D482E	Neutral	2.5	49	
F382L	Medium	2.1	41	
G390W	Medium	2.1	41	

^a Uniprot amino acid sequence numbering of primary isoform

^b MutationAssessor Score [216, 217], which predicts functional impact via conservation

Table 5.3: Expression yields for engineered clinical missense mutants of Src kinase domains with yields $\geq 2 \mu\text{g/mL}$ culture. Src kinase domain constructs with engineered clinical mutations with expression yields $\geq 2 \mu\text{g/mL}$ culture are listed, sorted by yield. Yield was determined by Caliper GX II quantitation of the expected size band and reported in $\mu\text{g/mL}$ culture, where total eluate volume was $80 \mu\text{L}$ purified from $900 \mu\text{L}$ bacterial culture. Wild-type (WT) controls for both Src and Abl (here, a single well for each) are shown as the first entry for each gene.

Src (254–536)	Mutation ^a	Functional Impact Score ^b	yield ($\mu\text{g/mL}$)	% of WT expression
WT	—	35.7	—	—
T456S	Neutral	80.9	227	
R388G	Medium	61.5	172	
K298E	High	54.5	153	
V380M	Neutral	51.7	145	
D368N	Neutral	49.9	140	
D521N	Low	42.8	120	
R463Q	Neutral	38.4	108	
R391C	Neutral	37.5	105	
E323D	Low	37.2	104	
A309V	Low	35.9	98	
G303D	Neutral	34.1	96	
R362Q	Neutral	33.6	94	
L361M	Medium	31.7	89	
A421V	Neutral	30.7	86	
V402L	Neutral	30.6	86	
V397M	Medium	29.8	84	
Q278E	Neutral	29.6	83	
Q312H	Low	29.5	83	
L353V	Medium	29.0	81	
L454V	Neutral	29.0	81	
P307R	Neutral	28.6	80	
V340I	Low	28.0	78	
P307S	Neutral	24.2	68	
D476N	Neutral	23.3	65	
D351N	Neutral	22.9	64	
T293A	Neutral	22.2	62	

Table 5.3 continued from previous page

Src (254–536)	Mutation ^a	Functional Impact Score ^b	yield (µg/mL)	% of WT expression
	S345C	Low	22.2	62
	P428S	Medium	22.2	62
	E507D	Neutral	20.7	58
	D389E	High	20.0	56
	R503Q	Neutral	17.3	49
	D407H	High	15.9	45
	R463L	Neutral	14.9	42
	G291C	Medium	11.9	33
	G347E	Medium	10.2	29
	R483W	High	9.8	27
	P487L	Medium	6.0	17
	R463W	Medium	5.2	15
	R362W	Low	3.9	11
	S493F	Low	3.0	8
	P491S	Low	2.2	6

^a Uniprot amino acid sequence numbering of primary isoform

^b MutationAssessor Score [216, 217], which predicts functional impact via conservation

Using our structural informatics pipeline, a database was built focusing on the kinases we found to be expressible in *E. coli*. To add the mutation data, we retrieved public datasets from cBioPortal [178, 179] along with annotations from Oncotator [218] through their respective web service APIs. We then added mutations and annotations from the MSKCC dataset [38] by extracting the mutations from a local copy of the dataset and retrieving annotations from Oncotator. The annotated mutations were filtered for mutations that occurred within the construct boundaries of our kinase domains. We found 63 unique clinical mutations appearing within our kinase domain construct boundaries for Abl and 61 for Src. We subsequently selected 48 mutants for Abl and 46 for Src to express, aiming for a panel of mutants distributed throughout the kinase domain (Figure 5.6A), with wild-type sequences included as controls. Mutations were introduced using site-directed mutagenesis and assayed for expression yields (Figure 5.6B). Those with yields above 2 μ g kinase/mL culture are listed in Tables 5.3 and 5.2.

High-expressing mutants appear to be distributed relatively uniformly throughout the kinase domain (Figure 5.6A). While the vast majority of the Src mutants expressed at a usable level, many of the Abl mutants expressed below the 2 μ g/mL threshold. This can primarily be attributed to the low level of expression for wild-type Abl construct (Table 5.1). In instances where kinase activity is not required, yield could be increased via the introduction of inactivating mutations [199] or further tailoring of expression and purification protocols.

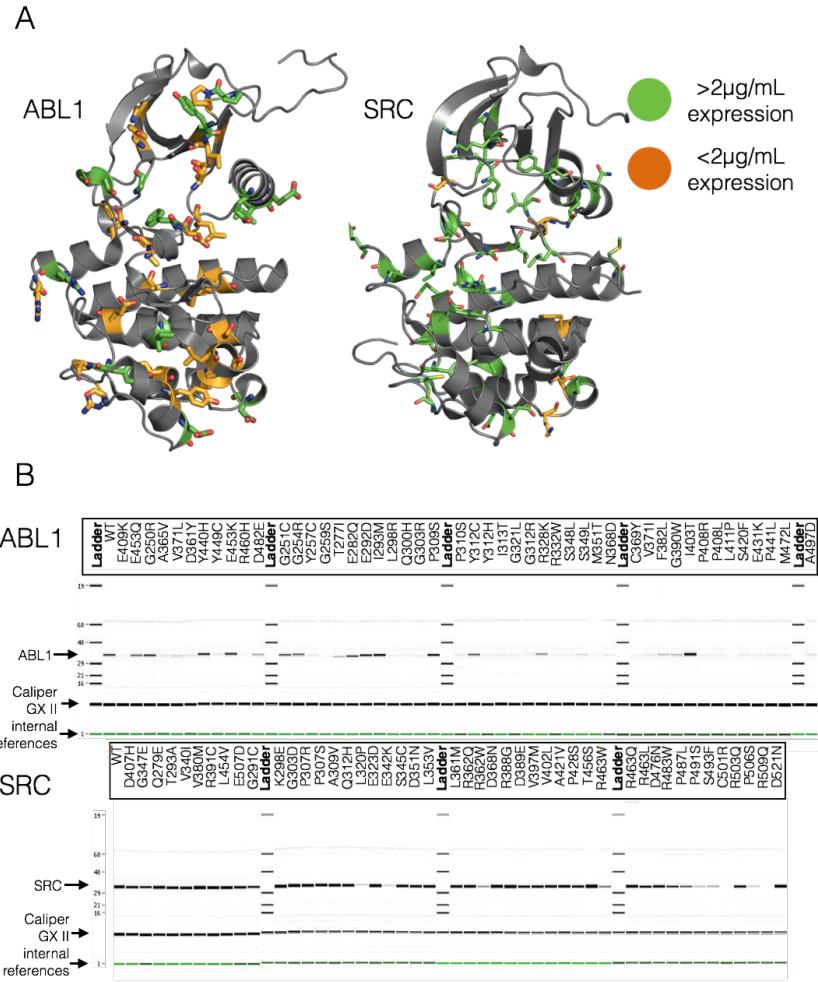


Figure 5.6: Expression yields for engineered clinically-derived Src and Abl missense mutants. (A) All Abl and Src clinically-identified mutants assessed in the expression screen are displayed as sticks. Mutants with expression yields $>2 \mu\text{g}/\text{mL}$ are colored green, while those with yields $<2 \mu\text{g}/\text{mL}$ are colored orange. Rendered structures are Abl (PDBID: 2E2B) and Src (PDBID: 4MXO) [213]. (B) Synthetic gel images showing ABL1 (top) or SRC (bottom) expression, with wells labeled by missense mutation. Yield was determined by Caliper GX II quantitation of the expected size band and reported in $\mu\text{g}/\text{mL}$ culture, where total eluate volume was $120 \mu\text{L}$ following nickel bead pulldown purification from $900 \mu\text{L}$ bacterial culture. Residue mutations use numbering for the Uniprot canonical isoform.

5.4 Methods

5.4.1 Semi-automated selection of kinase construct sequences for *E. coli* expression

Selection of human protein kinase domain targets

Human protein kinases were selected by querying the UniProt API (query date 30 May 2014) for any human protein with a domain containing the string "protein kinase", and which was manually annotated and reviewed (i.e. a Swiss-Prot entry). The query string used was:

```
taxonomy:"Homo sapiens (Human) [9606]" AND domain:"protein kinase" AND reviewed:yes
```

Data was returned by the UniProt API in XML format and contained protein sequences and relevant PDB structures, along with many other types of genomic and functional information. To select active protein kinase domains, the UniProt domain annotations were searched using the regular expression ^Protein kinase(?!; truncated)(?!; inactive), which excludes certain domains annotated "Protein kinase; truncated" and "Protein kinase; inactive". Sequences for the selected domains, derived from the canonical isoform as determined by UniProt, were then stored.

Matching target sequences with relevant PDB constructs

Each target kinase gene was matched with the homologous in any other species, if present, and all UniProt data was downloaded. This data included a list of PDB

structures which contain the protein, and their sequence spans in the coordinates of the UniProt canonical isoform. PDB structures which did not include the protein kinase domain or truncated more than 30 residues at each end were filtered out. PDB coordinate files were then downloaded for each remaining PDB entry. The coordinate files contain various metadata, including the EXPRESSION_SYSTEM annotation, which was used to filter PDB entries for those which include the phrase "ESCHERICHIA COLI". The majority of PDB entries returned had an EXPRESSION_SYSTEM tag of "ESCHERICHIA COLI", while a small number had "ESCHERICHIA COLI BL21" or "ESCHERICHIA COLI BL21(DE3)".

The PDB coordinate files also contain SEQRES records, which should contain the protein sequence used in the crystallography or NMR experiment. According to the PDB-101 (<http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/primary-sequences-and-the-pdb-format>), the SEQRES should include the "sequence of each chain of linear, covalently-linked standard or modified amino acids or nucleotides. It may also include other residues that are linked to the standard backbone in the polymer." However, we found that these records are very often misannotated, instead representing only the crystallographically resolved residues. Since expression levels can be greatly affected by insertions or deletions of only one or a few residues at either terminus [219], it is important to know the full experimental sequence. To measure the authenticity of a given SEQRES record, we developed a simple metric by hypothesizing that most crystal structures would likely have at least one or more unresolved residues at one or both termini and that the presence of an expression tag, which is typically not crystallographically resolved, would indicate an authentic SEQRES record. To achieve this, unresolved residues were first defined by comparing the SEQRES sequence to the resolved sequence, using the SIFTS service to determine which

residues were not present in the canonical isoform sequence [220]. Regular expression pattern matching was used to detect common expression tags at the N- or C-termini. Sequences with a detected expression tag were given a score of 2, while those with any unresolved sequence at the termini were given a score of 1, and the remainder were given a score of 0. This data was stored to allow for subsequent selection of PDB constructs based on likely authenticity in later steps. The number of residues extraneous to the target kinase domain, and the number of residue conflicts with the UniProt canonical isoform within that domain span were also stored for each PDB sequence.

Plasmid libraries

As a source of kinase DNA sequences for subcloning, we purchased three kinase plasmid libraries: the [Addgene Human Kinase ORF kit](#), a kinase library from the Structural Genomics Consortium (SGC), Oxford (<http://www.thesgc.org>), and a kinase library from the [PlasmID Repository](#) maintained by the Dana-Farber/Harvard Cancer Center. Annotated data for the kinases in each library was used to match them to the human protein kinases selected for this project. The plasmid open reading frames (ORFs) were translated into protein sequences and aligned against the target kinase domain sequences from UniProt. Also calculated were the number of extraneous protein residues in the ORF, relative to the target kinase domain sequence, and the number of residue conflicts with the UniProt sequence. Our aim was to subclone the chosen sequence constructs from these library plasmids into our expression plasmids.

Selection of sequence constructs for expression

Of the kinase domain targets selected from UniProt, we filtered out those with no matching plasmids in our available plasmid libraries or no suitable PDB construct sequences. For this purpose, a suitable PDB construct sequence was defined as any with an authenticity score greater than zero (see above). Library plasmid sequences and PDB constructs were aligned against each Uniprot target domain sequence, and various approaches were considered for selecting the construct boundaries to use for each target, and the library plasmid to subclone it from. Candidate construct boundaries were drawn from two sources: PDB constructs and the SGC plasmid library, has been successfully tested for *E. coli* expression.

For most of the kinase domain targets, multiple candidate construct boundaries were available. To select the most appropriate construct boundaries, we sorted them first by authenticity score, then by the number of conflicts relative to the UniProt domain sequence, then by the number of residues extraneous to the UniProt domain sequence span. The top-ranked construct was then chosen. In cases where multiple library plasmids were available, these were sorted first by the number of conflicts relative to the UniProt domain sequence, then by the number of residues extraneous to the UniProt domain sequence span, and the top-ranked plasmid was chosen. This process resulted in a set of 96 kinase domain constructs, which (by serendipity) matched the 96-well plate format we planned to use for parallel expression testing. We selected these constructs for expression testing.

An interactive table of the selected plasmids, constructs, and aligned PDB files can be viewed at <http://choderlab.org/kinome-expression>.

Automation of the construct selection process

While much of this process was performed programmatically, many steps required manual supervision and intervention to correct for exceptional cases. While these exceptions were encoded programmatically as overrides to ensure the scheme could be reproduced from existing data, we hope to eventually develop a fully automated software package for the selection of expression construct sequences for a given protein family, but this was not possible within the scope of this work.

5.4.2 Mutagenesis protocol

Point mutations were introduced with a single-primer QuikChange reaction. Primers were designed to anneal at 55°C both upstream and downstream of the point mutation, and with a total length of approximately 40 bases. At the codon to be modified, the fewest possible number of bases was changed. Plasmid template (160 ng) was mixed with 1 μ M primer in 1x PfuUltra reaction buffer, with 0.8 mM dNTPs (0.2 mM each) and 1 U PfuUltra High-Fidelity DNA polymerase (Agilent), in a total volume of 20 μ L. Thermocycler settings were 2 min at 95°C, followed by 18 cycles of 20s at 95°C, 1 min at 53°C, 12 min at 68°C (2min/kb), then 1 minute at 68°C. After cooling to room temperature, 4 μ L of the PCR reaction was added to 16 μ L CutSmart Buffer (NEB) containing 10 U DpnI (NEB). After incubation for 2.5 hours at 37°C, 6 μ L of this mixture was used to directly transform XL1-Blue chemically competent cells (Agilent) according to the manufacturer's protocol. Transformants were picked for plasmid mini-preps and the presence of the point mutations was confirmed by sequencing.

5.4.3 Expression testing

For each target, the selected construct sequence was subcloned from the selected DNA plasmid. Expression testing was performed at the QB3 MacroLab (QB3 MacroLab, University of California, Berkeley, CA 94720)

[<http://qb3.berkeley.edu/macrolab/>], a core facility offering automated gene cloning and recombinant protein expression and purification services.

Each kinase domain was tagged with a N-terminal His10-TEV and coexpressed with either the truncated YopH164 for Tyr kinases or lambda phosphatase for Ser/Thr kinases. All construct sequences were cloned into the 2BT10 plasmid, an AMP resistant ColE1 plasmid with a T7 promoter, using ligation-independent cloning (LIC). The inserts were generated by PCR using the LICv1 forward (TACTTCCAATCCAATGCA) and reverse (TTATCCACTTCCAATGTTATT) tags on the primers. Gel purified PCR products were LIC treated with dCTP. Plasmid was linearized, gel purified, and LIC-treated with dGTP. LIC-treated plasmid and insert were mixed together and transformed into XL1-Blues for plasmid preps.

Expression was performed in Rosetta2 cells (Novagen) grown with Magic Media (Invitrogen autoinducing medium), 100 µg/mL of carbenicillin and 100 µg/mL of spectinomycin. Single colonies of transformants were cultivated with 900 µL of MagicMedia into a gas permeable sealed 96-well block. The cultures were incubated at 37°C for 4 hours and then at 16°C for 40 hours while shaking. Next, cells were centrifuged and the pellets were frozen at -80°C overnight. Cells were lysed on a rotating platform at room temperature for an hour using 700 µL of SoluLyse (Genlantis) supplemented with 400 mM NaCl, 20 mM imidazole, 1 µg/mL pepstatin, 1 µg/mL leupeptin and 0.5 mM PMSF.

For protein purification, 500 µL of the soluble lysate was added to a 25 µL Protino

Ni-NTA (Machery-Nagel) agarose resin in a 96-well filter plate. Nickel Buffer A (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl, 20 mM imidazole, 1 mM BME) was added and the plate was shaken for 30 min at room temperature. The resin was washed with 2 mL of Nickel Buffer A. For the 96-kinase expression experiment, target proteins were eluted by a 2 hour incubation at room temperature with 10 μ g of TEV protease in 80 μ L of Nickel Buffer A per well and a subsequent wash with 40 μ L of Nickel Buffer A to maximize protein release. Nickel Buffer B (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl, 400 mM imidazole, 1 mM BME) was used to elute TEV resistant material remaining on the resin. Untagged protein eluted with TEV protease was run on a LabChip GX II Microfluidic system to analyze the major protein species present.

For the clinical mutant and Abl1 construct boundaries expression experiments, target proteins were washed three times with Nickel Buffer A prior to elution in 80 μ L Nickel Buffer B. The eluted protein was run on a LabChip GX II Microfluidic system to analyze with major protein species were present.

5.4.4 Fluorescence-based thermostability assay

To assess whether the highly-expressed wild-type kinase constructs are folded, a thermofluor thermostability assay [208–210] was performed for kinase constructs that have a minimum of 0.24 mg/mL protein concentration in the eluate. After diluting 9 μ L of eluate by 1 μ L dye, the effective assay concentration is 0.216 mg/mL minimum in 10 μ L assay volume. Previous optimization efforts in the lab determined that 0.20 mg/mL was the lower limit of well-defined T_m detection. This minimum concentration also ensured that the kinase was present at roughly an order of magnitude concentration higher than contaminating TEV

protease.

Kinase expression panel eluates, which were kept in 96-well deep well plate frozen at -80°C for 2 years prior to the thermal stability assay, were thawed in an ice-water bath for 30 min. 9 μ L of each kinase eluate was added to a 384 well PCR plate (4titude-0381). 100X SYPRO Orange dye solution was prepared from a 5000X DMSO solution of SYPRO Orange Protein Gel Stain (Life Technologies, Ref S6650, LOT 1790705) by dilution in distilled water. In initial experiments, SYPRO Orange dye solution was diluted in kinase binding assay buffer (20 mM Tris 0.5 mM TCEP pH 8), which caused the dye to precipitate out of solution. Particulates in the dye solution were pelleted by tabletop centrifugation (2 min, 5000 RCF) and the solution was kept covered with aluminum foil in the dark to prevent photodamage. 1 μ L of 100X dye solution was added to each kinase eluate sample in 384-well PCR plate. The plate was sealed with Axygen UC-500 Ultra Clear Pressure Sensitive sealing film. To remove any air bubbles, the sample plate was centrifuged for 30 sec with 250 g using Bionex HiG4 centrifuge. Sample mixing was performed by orbital shaking with Inheco shakers for 2 min at 1000 RPM.

A thermofluor melt was performed using a LightCycler 480 (Roche) qPCR instrument using an excitation filter of 465 nm (half bandwidth 25 nm) and emission filter at 580 nm (half bandwidth 20 nm). LightCycler 480 Software Version 1.5.1 was used to operate the instrument and analyze the results. The temperature was held at 25°C for 15 s before ramping up to 95°C with a ramp rate of 0.06°C/s. During temperature ramp 10 fluorescence acquisitions/°C were recorded with dynamic integration time mode, melt factor of 1, quant factor of 10, and maximum integration time of 2 sec. Thermal protein denaturation causes hydrophobic patches of protein to be exposed, which SYPRO Orange dye can

bind. Binding of SYPRO Orange dye is detected as an increase in fluorescence at 580 nm. Presence of a clear thermal denaturation peak in the absolute value of the derivative of the fluorescence as a function of temperature serves as an indication that the proteins were well-folded. Observed fluorescence was plotted as a function of temperature, and a melting temperature T_m was determined as the maximum of the absolute value of its first derivative.

5.4.5 ATP-competitive inhibitor binding fluorescence assay

To determine whether the expressed kinases had a well-folded ATP-binding site, we assessed whether the eluted kinase was capable of binding the ATP-competitive small molecule kinase inhibitor bosutinib. We designed fluorescence-based binding assays following earlier work reporting that this quinoline-scaffold inhibitor undergoes a strong increase in fluorescence upon binding (even weakly) to kinase ATP-binding sites [212]. By titrating in the ligand to close to the solubility limit, even weak binding to the ATP-binding site can be detected by observing emission increases around 450 nm during excitation at 280 nm.

For 33 of the kinases in our expression panel, 0.5 μ M kinase solutions from kinase expression panel eluates were prepared in kinase binding assay buffer (20 mM Tris 0.5 mM TCEP pH 8) for a final volume of 100 μ L in a black 96-well vision plate (4titude-0223). Six low-expressing kinases (Figure 5.5, panels 39-44) were prepared by diluting 20 μ L of eluate in kinase binding assay buffer (20 mM Tris 0.5 mM TCEP pH 8) to a final volume of 100 μ L, for final concentrations below 0.5 μ M. The plate was shaken for 2 min clockwise and 2 min counter-clockwise by orbital shaking with Inheco shakers at 2000 RPM and centrifuged for 30 sec with

1000 g using Bionex HiG4 centrifuge. Fluorescence emission spectra were measured from 370 nm to 600 nm (20 nm bandwidth) in 5 nm steps using 280 nm excitation (10 nm bandwidth) from both the top and bottom of the well using a Tecan Infinite M1000 PRO.

Bosutinib free base (LC Labs, cat no. B-1788, lot no. BSB-103, M.W. 530.45 Da) was dispensed directly from a roughly 10 mM DMSO stock solution to the assay solution using a Tecan HP D300 Digital Dispenser. The 10 mM DMSO stock solution was prepared gravimetrically using an automated balance (Mettler Toledo Balance XPE205 with LabX Laboratory Software) by dispensing 39.02 mg solid Bosutinib powder stored under nitrogen gas at 25°C into 8.0499 g DMSO (Alfa Aesar, cat no. 42780, log no. Y25B604, density 1.1004 g/mL at ambient temperature) which is kept dry under argon gas at 25°C. To minimize atmospheric water absorption due to the hygroscopic nature of DMSO, the 10 mM stock solution was pipetted into wells of a 96-well stock plate by an automated liquid handling device (Tecan EVO 200 with air LiHa) and sealed with foil seal (PlateLoc). Ligand was dispensed to the assay plate with HP D300 (using aliquots of stock solution pipetted from a freshly pierced stock plate well) targeting a roughly geometrically-increasing series of ligand concentrations in each well to achieve the following total ligand concentrations after each dispense: 0.008 μ M, 0.013 μ M, 0.023 μ M, 0.038 μ M, 0.064 μ M, 0.109 μ M, 0.183 μ M, 0.308 μ M, 0.519 μ M, 0.875 μ M, 1.474 μ M, 3.174 μ M, 6.037 μ M, 10.862 μ M, 18.991 μ M. The plate was shaken by HP D300 for 10 sec after usage of each dispensehead. After each titration, the plate was shaken with Inheco shakers (2 min clockwise and counter-clockwise, 2000 RPM, orbital shaking) and centrifuged (30 sec, 1000 g) using a Bionex HiG4 centrifuge. Fluorescence spectra from 370 nm to 600 nm (bandwidth 20 nm) in 5 nm steps using 280 nm excitation (bandwidth 10 nm) were read from both the top and bottom of the well using a Tecan Infinite M1000 PRO.

In total, the experiment took 17.5 hours to complete due to the time-consuming spectral read after each dispense, likely resulting in significant evaporation from some wells during the experiment.

ATP-competitive binding was analyzed qualitatively for each kinase by plotting the fluorescence spectra as a function of concentration to detect concentration-dependent increases in fluorescence. As a control for background ligand fluorescence independent of protein binding, fluorescence spectra of three replicates of ligand into buffer titrations were plotted. As a positive control, MK14 produced by a validated large scale expression protocol (see Supplementary Methods) from the same plasmid used in the high-throughput protocol was included. To control for non-specific binding to unfolded protein, we included boiled MK14 (prepared from the large scale expression of MK14 by boiling at 95°C for 10 min). A concentration-dependent increase in fluorescence was interpreted as evidence that the ATP-binding site of the kinase was well folded and allowed for bosutinib binding. Due to the length of the experiment, it is possible that evaporation reduced the well volume below 100 μ L and potentially caused bosutinib to reach higher concentration levels than expected. This creates uncertainty for data points, as bosutinib may either be a higher concentration (due to evaporation) or a lower concentration (due to potential precipitation caused by lower well volumes) than expected. For this reason, we have interpreted the experiment as qualitative evidence of binding, instead of quantitatively. Bosutinib binding is an indication of proper folding of the ATP binding pocket of these recombinantly expressed kinase constructs.

5.4.6 Large Scale expression and purification protocol for MK14

Large scale expression of MK14 was performed at the QB3 MacroLab (QB3 MacroLab, University of California, Berkeley, CA 94720

[<http://qb3.berkeley.edu/macrolab/>], a core facility offering automated gene cloning and recombinant protein expression and purification services.

Rosetta2(DE3)pLysS cells (Novagen) were used to co-express MK14 (same plasmid as from the high-throughput kinase expression panel) and 13SA Lamda phosphatase. The cells were grown in 2YT Medium (16 g/L Tryptone, 10 g/L Yeast Extract, 5 g/L NaCl) to OD600 of 0.5 at 37°C. The culture was cooled to 16°C and induced with 0.5 mM IPTG overnight. The cultures were pelleted at 5000 rpm for 30 min and resuspended in 20 mL Nickel buffer A (25 mM HEPES pH 7.5, 10% glycerol, 400mM NaCl, 20 mM imidazole, 5 mM BME) with the following protease inhibitors: 1 µg/mL leupeptin, 1 µg/mL pepstatin, and 0.5 mM PMSF). The resuspended cells were frozen at -80°C.

When ready for purification, the cells were thawed and ruptured using a homogenizer (Avestin C3, 15000psi, 3 passes). The broken cells were pelleted at 15000 rpm for 30 min (SS34 rotor). Clarified lysate was loaded onto a 5 mL HisTrap FF Crude column (GE Healthcare) and washed with Nickel buffer A to remove any unbound material. The protein was eluted with Nickel buffer B (25 mM HEPES pH 7.5, 10% glycerol, 400mM NaCl, 400 mM imidazole, 5 mM BME) and pooled for buffer exchange into Nickel buffer A on a HiPrep 26/10 Desalting Column (GE Healthcare). Rough protein yields were quantified using theoretical extinction coefficients calculated using ProtParam (<http://ca.expasy.org/tools/protparam.html>). The His tag was cleaved off of MK14 by incubation with TEV protease (25°C, 2 hours, 1:20 mass ratio).

After tag cleavage, the sample was run over a 5 mL HisTrap FF Crude column (GE Healthcare) with Nickel buffer A. The flow-through was collected, concentrated to roughly 5mL using centrifugal concentrators (10 kDa MWCO, Millipore) and loaded onto a HiPrep 16/60 Sephadryl S-200 HR column (GE Healthcare). The sample was equilibrated into Gel Filtration buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 5% glycerol, 1 mM DTT) and fractions containing MK14 were pooled and concentrated (10 kDa MWCO centrifugal concentrators, Millipore). 500 μ L aliquots of MK14 were snap frozen in liquid nitrogen and stored at -80°C. Quantification by theoretical extinction coefficient suggests the final MK14 concentration was roughly 4.0 mg/mL (97 μ M), roughly 22.4 mg/L of culture yield.

5.5 Discussion

We have demonstrated that a simple, uniform, automatable protocol is able to achieve useful bacterial expression yields for a variety of kinase domain constructs. While yields could likely be further improved by a variety of methods—such as the addition of solubility-promoting tags, construct domain boundary and codon optimization, or mutations to improve the solubility or ablate catalytic activity—the simplicity of this approach suggests widespread utility of automated bacterial expression for biophysical, biochemical, and structural biology work for the further study of human kinase domains.

Our expression test of 81 different construct boundaries of the Abl kinase domain demonstrated a surprising sensitivity of expression yields to the precise choice of boundary. This sensitivity may be related to where the construct is truncated with respect to the secondary structure of the protein, as disrupting secondary

structure could cause the protein to improperly fold, leading to low soluble protein yield even when total expression is high. Of note, the highest expressing C-terminal boundaries for Abl were residues 511 and 512. These residues fall in the regulatory alpha helix α I [20]. This helix has been shown to undergo a dramatic conformational change upon binding to the myristoylated N-terminal cap, which introduces a sharp “kink” in residues 516–519. These residues may lead to higher levels of soluble expression by truncating an secondary structural element that is unusually flexible. Indeed, this helix is not resolved in some X-ray structures (PDBID:2E2B) [203], further suggesting that this helix is less thermodynamically stable than expected. Control replicates of three constructs indicate good repeatability of expression yields in the high-throughput format. This screen suggests that optimization of construct boundaries could potentially further greatly increase yields of poorly expressing kinase domains. Codon optimization for bacterial expression could also increase expression for kinase domains with low yield due to codon bias [221], as could coexpression with chaperones [222].

For those kinases that did express, a fluorescence-based thermostability assay indicated that many of the highest-expressing kinases are well folded. An ATP-competitive inhibitor binding fluorescent assay provides qualitative evidence that the 39 kinases that had sufficiently high expression levels to be assayed have a well-formed ATP-binding site capable of binding bosutinib, a small molecule ATP-competitive kinase inhibitor. Taken together, these two experiments demonstrate that our expression protocol produces folded kinases with utility for biophysical experiments and drug design.

The tolerance of these bacterial constructs to many engineered clinical missense mutations suggests a promising route to the high-throughput biophysical

characterization of the effect of clinical mutants on anticancer therapeutics. Mutations that did not express well may destabilize the protein, or may increase the specific activity of the kinase. A higher specific activity would require more phosphatase activity, wasting ATP to prevent high levels of phosphorylation that have been hypothesized to cause difficulty expressing kinases without a coexpressed phosphatase in bacteria [199]. Mutations that are destabilizing may show improved expression if coexpressed with more elaborate chaperones such as GroEL and Trigger factor [222]. Mutations that increase the specific activity of the kinase might also express better when combined with an inactivating mutation.

High-throughput automated kinase expression could be combined with enzymatic or biophysical techniques for characterizing the potency of a variety of clinical kinase inhibitors to assess which mutations confer resistance or sensitivity. While the process of engineering, expressing, purifying, and assaying mutants currently takes approximately two weeks, it is possible that new techniques for cell-free bacterial expression [223, 224] may reduce this time to a matter of days or hours in a manner that might be compatible with clinical time frames to impact therapeutic decision-making.

We hope that other laboratories find these resources useful in their own work.

5.6 Author Contributions

Conceptualization, JDC, DLP, SKA, MI, LRL, SMH, NML, MAS; Methodology, DLP, MI, LRL, SMH, SKA, JDC, NML, MAS; Software, DLP, JDC, SMH; Formal Analysis, SKA, JDC, MI, SMH; Investigation, MI, LRL, SG, CJ, SKA, SMH; Resources, CJ, SG; Data Curation, SKA, MI, LRL, DLP, JMB; Writing-Original

Draft, SKA, LRL, DLP, JDC, SG, SMH, MI; Writing - Review and Editing, SKA, JDC, MI, LRL, SHM, SG, CJ, NML, MAS; Visualization, SKA, JDC, MI, SMH; Supervision, JDC, NML, MAS; Project Administration, SKA, JDC, MI, SMH; Funding Acquisition, JDC, SMH

5.7 Acknowledgments

DLP, SMH, LRL, SKA, MI, and JDC acknowledge support from the Sloan Kettering Institute. This work was funded in part by the Marie-Josée and Henry R. Kravis Center for Molecular Oncology, the National Institutes of Health (NIH grant R01 GM121505 and National Cancer Institute Cancer Center Core grant P30 CA008748), the Functional Genomics Institute (FGI) at MSKCC, and a Louis V. Gerstner Young Investigator Award. MAS acknowledges funding support by NIH grant R35 GM119437. The authors are grateful to Gregory Ross (MSKCC) for assistance in preparing the computational infrastructure for selecting clinical point mutants, and to Sarah E. Boyce (current address: Schrödinger, New York, NY) for assistance with multiple stages of this project. We gratefully acknowledge the members of the MSKCC Molecular Diagnostics Service in the Department of Pathology for their efforts in collecting and compiling mutations for Abl and Src kinases used here. We thank the Kuriyan lab for the gift of pCDFDuet1-YOPH plasmid. The authors are grateful to [Addgene](#) for their help in making the plasmids generated by this work available to the research community at minimal cost.

BIBLIOGRAPHY

- [1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: a cancer journal for clinicians. 2018 Jan; 68(1):7–30.
- [2] Cohen P, Tcherpakov M. Will the ubiquitin system furnish as many drug targets as protein kinases? Cell. 2010 Nov; 143(5):686–693.
- [3] Robert Roskoski Jr. USFDA Approved Protein Kinase Inhibitors. . 2017; <http://www.brimr.org/PKI/PKIs.htm>, updated 3 May 2017.
- [4] Paul, Steven M, Mytelka, Daniel S, Dunwiddie, Christopher T, Persinger, Charles C, Munos, Bernard H, Lindborg, Stacy R, Schacht, Aaron L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery. 2010 Mar; 9(3):203–214.
- [5] Sanchez-Vega F, Hechtman JF, Castel P, Ku GY, Tuvy Y, Won H, Fong CJ, Bouvier N, Nanjangud GJ, Soong J, Vakiani E, Schattner M, Kelsen DP, Lefkowitz RA, Brown K, Lacouture ME, Capanu M, Mattar M, Qeriqi B, Cecchi F, et al. EGFR and MET Amplifications Determine Response to HER2 Inhibition in ERBB2-Amplified Esophagogastric Cancer. Cancer discovery. 2018 Nov; .
- [6] Bose R, Kavuri SM, Searleman AC, Shen W, Shen D, Koboldt DC, Monsey J, Goel N, Aronson AB, Li S, Ma CX, Ding L, Mardis ER, Ellis MJ. Activating HER2 mutations in HER2 gene amplification negative breast cancer. Cancer discovery. 2013 Feb; 3(2):224–237.
- [7] Prahallas A, Sun C, Huang S, Di Nicolantonio F, Salazar R, Zecchin D, Beijersbergen RL, Bardelli A, Bernards R. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature. 2012 Mar; 483(7387):100–103.
- [8] Engelman JA, Zejnnullahu K, Mitsudomi T, Song Y, Hyland C, Park JO, Lindeman N, Gale CM, Zhao X, Christensen J, Kosaka T, Holmes AJ, Rogers AM, Cappuzzo F, Mok T, Lee C, Johnson BE, Cantley LC, Jänne PA. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science. 2007 May; 316(5827):1039–1043.
- [9] Pao W, Miller VA, Politi KA, Riely GJ, Somwar R, Zakowski MG Maureen Fnd Kris, Varmus H. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS medicine. 2005 Mar; 2(3):e73.
- [10] Drilon A, Nagasubramanian R, Blake JF, Ku N, Tuch BB, Ebata K, Smith S, Lauriault V, Kolakowski GR, Brandhuber BJ, Larsen PD, Bouhana KS, Winski SL, Hamor R, Wu WI, Parker A, Morales TH, Sullivan FX, DeWolf WE, Wollenberg LA, et al. A Next-Generation TRK Kinase Inhibitor Overcomes Acquired Resistance to Prior TRK Kinase Inhibition in Patients with TRK Fusion-Positive Solid Tumors. Cancer discovery. 2017 Sep; 7(9):963–972.
- [11] Knight ZA, Lin H, Shokat KM. Targeting the Cancer Kinome through Polypharmacology. Nat Rev Cancer. 2010; 10(2):130.
- [12] Chandarlapaty S, Sawai A, Scaltriti M, Rodrik-Outmezguine V, Grbovic-Huezo O, Serra V, Majumder PK, Baselga J, Rosen N. AKT Inhibition Relieves Feedback Suppression of Receptor Tyrosine Kinase Expression and Activity. Cancer Cell. 2011 Jan; 19(1):58–71. doi: 10.1016/j.ccr.2010.10.031.

- [13] Rudmann DG. On-target and off-target-based toxicologic effects. *Toxicol Pathol.* 2013 Feb; 41(2):310–314.
- [14] Liu S, Kurzrock R. Toxicity of targeted therapy: Implications for response and impact of genetic polymorphisms. *Cancer Treat Rev.* 2014 Aug; 40(7):883–891.
- [15] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The Protein Kinase Complement of the Human Genome. *Science.* 2002 Dec; 298(5600):1912–1934.
- [16] Cowan-Jacob SW, Fendrich G, Floersheimer A, Furet P, Liebetanz J, Rummel G, Rheinberger P, Centeleghe M, Fabbro D, Manley PW, IUCr. Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta Crystallogr D Biol Crystallogr.* 2007 Jan; 63(1):80–93.
- [17] Seeliger MA, Nagar B, Frank F, Cao X, Henderson MN, Kuriyan J. c-Src Binds to the Cancer Drug Imatinib with an Inactive Abl/c-Kit Conformation and a Distributed Thermodynamic Penalty. *Structure.* 2007 Mar; 15(3):299–311.
- [18] Huse M, Kuriyan J. The conformational plasticity of protein kinases. *Cell.* 2002 Jan; 109(3):275–282.
- [19] Harrison SC. Variation on an Src-like theme. *Cell.* 2003 Mar; 112(6):737–740.
- [20] Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J. Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell.* 2003 Mar; 112(6):859–871.
- [21] Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP. Comprehensive Analysis of Kinase Inhibitor Selectivity. *Nat Biotechnol.* 2011 Oct; 29(11):1046–1051. doi: [10.1038/nbt.1990](https://doi.org/10.1038/nbt.1990).
- [22] Apsel B, Blair JA, Gonzalez B, Nazif TM, Feldman ME, Aizenstein B, Hoffman R, Williams RL, Shokat KM, Knight ZA. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat Chem Biol.* 2008 Nov; 4(11):691–699.
- [23] Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? *Curr Opin Struct Biol.* 2006 Feb; 16(1):127–136.
- [24] Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol.* 2008 Nov; 4(11):682–690.
- [25] Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer.* 2009 Jan; 9(1):28–39.
- [26] Huggins DJ, Sherman W, Tidor B. Rational approaches to improving selectivity in drug design. *J Med Chem.* 2012 Feb; 55(4):1424–1444.
- [27] Mendoza MC, Er EE, Blenis J. The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. *Trends Biochem Sci.* 2011 Jun; 36(6):320–328.
- [28] Tricker EM, Xu C, Uddin S, Capelletti M, Ercan D, Ogino A, Pratillas CA, Rosen N, Gray NS, Wong KK, Jänne PA. Combined EGFR/MEK Inhibition Prevents the Emergence of Resistance in EGFR-Mutant Lung Cancer. *Cancer Discov.* 2015 Sep; 5(9):960–971.

- [29] Chandarlapaty S, Sawai A, Scaltriti M, Rodrik-Outmezguine V, Grbovic-Huezo O, Serra V, Majumder PK, Baselga J, Rosen N. AKT Inhibition Relieves Feedback Suppression of Receptor Tyrosine Kinase Expression and Activity. *Cancer Cell*. 2011 Jan; 19(1):58–71.
- [30] Bailey ST, Zhou B, Damrauer JS, Krishnan B, Wilson HL, Smith AM, Li M, Yeh JJ, Kim WY. mTOR Inhibition Induces Compensatory, Therapeutically Targetable MEK Activation in Renal Cell Carcinoma. *PLoS One*. 2014 Sep; 9(9):e104413.
- [31] Littlefield P, Moasser MM, Jura N. RETRACTED: An ATP-Competitive Inhibitor Modulates the Allosteric Function of the HER3 Pseudokinase. *Chem Biol*. 2014 Apr; 21(4):453–458. doi: [10.1016/j.chembiol.2014.02.011](https://doi.org/10.1016/j.chembiol.2014.02.011).
- [32] Yun, Cai-Hong, Boggon, Titus J, Li, Yiqun, Woo, Michele S, Greulich, Heidi, Meyerson, Matthew, Eck, Michael J. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell*. 2007 Mar; 11(3):217–227.
- [33] Gajiwala KS, Feng J, Ferre R, Ryan K, Brodsky O, Weinrich S, Kath JC, Stewart A. Insights into the Aberrant Activity of Mutant EGFR Kinase Domain and Drug Recognition. *Structure*. 2013; 21(2):209–219.
- [34] Jia Y, Yun CH, Park E, Ercan D, Manuia M, Juarez J, Xu C, Rhee K, Chen T, Zhang H, Palakurthi S, Jang J, Lelais G, DiDonato M, Bursulaya B, Michelllys PY, Epple R, Marsilje TH, McNeill M, Lu W, et al. Overcoming EGFR(T790M) and EGFR(C797S) resistance with mutant-selective allosteric inhibitors. *Nature*. 2016 Jun; 534(7605):129–132.
- [35] Politi K, Ayeni D, Lynch T. The Next Wave of EGFR Tyrosine Kinase Inhibitors Enter the Clinic. *Cancer Cell*. 2015 Jun; 27(6):751–753.
- [36] Song Z, Wang M, Zhang A. Alectinib: a novel second generation anaplastic lymphoma kinase (ALK) inhibitor for overcoming clinically-acquired resistance. *Acta Pharmaceutica Sinica B*. 2015 Jan; 5(1):34–37.
- [37] Varghese AM, Berger MF. Advancing clinical oncology through genome biology and technology. *Genome biology*. 2014 Aug; 15(8):427.
- [38] Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM, Hellmann MD, Barron DA, Schram AM, Hameed M, Dogan S, Ross DS, Hechtman JF, DeLair DF, Yao J, Mandelker DL, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*. 2017 May; 159:676.
- [39] Garraway LA. Genomics-driven oncology: framework for an emerging paradigm. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2013 May; 31(15):1806–1814.
- [40] Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. *BMC biology*. 2011; 9(1):71.
- [41] Lin JH, Perryman AL, Schames JR, McCammon JA. The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers*. 2003 Jan; 68(1):47–62.

- [42] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*. 2004 Nov; 3(11):935–949.
- [43] Amaro RE, Baron R, McCammon JA. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of computer-aided molecular design*. 2008 Jan; 22(9):693–705.
- [44] Hazuda DJ, Anthony NJ, Gomez RP, Jolly SM, Wai JS, Zhuang L, Fisher TE, Embrey M, Guare JP, Egbertson MS, Vacca JP, Huff JR, Felock PJ, Witmer MV, Stillmok KA, Danovich R, Grobler J, Miller MD, Espeseth AS, Jin L, et al. A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. *Proceedings of the National Academy of Sciences*. 2004 Aug; 101(31):11233–11238.
- [45] Norman RA, Toader D, Ferguson AD. Structural approaches to obtain kinase selectivity. *Trends in pharmacological sciences*. 2012 May; 33(5):273–278.
- [46] Shoichet BK. Virtual screening of chemical libraries. *Nature*. 2004 Dec; 432(7019):862–865.
- [47] Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A Critical Assessment of Docking Programs and Scoring Functions. *Journal of medicinal chemistry*. 2006 Oct; 49(20):5912–5931.
- [48] Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical free energy methods for drug discovery: progress and challenges. *Curr Opin Struct Biol*. 2011 Apr; 21(2):150–160.
- [49] Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical science*. 2016 Jan; 7(1):207–218.
- [50] Mikulskis P, Genheden S, Ryde U. A large-scale test of free-energy simulation estimates of protein-ligand binding affinities. *Journal of chemical information and modeling*. 2014 Oct; 54(10):2794–2806.
- [51] Huang J, MacKerell AD. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J Comput Chem*. 2013 Sep; 34(25):2135–2145. doi: [10.1002/jcc.23354](https://doi.org/10.1002/jcc.23354).
- [52] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput*. 2015 Aug; 11(8):3696–3713. doi: [10.1021/acs.jctc.5b00255](https://doi.org/10.1021/acs.jctc.5b00255).
- [53] Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyán D, Dahlgren MK, Knight JL, Kaus JW, Cerutti DS, Krilov G, Jorgensen WL, Abel R, Friesner RA. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J Chem Theory Comput*. 2016 Jan; 12(1):281–296. doi: [10.1021/acs.jctc.5b00864](https://doi.org/10.1021/acs.jctc.5b00864).
- [54] Cournia Z, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of chemical information and modeling*. 2017 Dec; 57(12):2911–2937.

- [55] Brown SP, Muchmore SW, Hajduk PJ. Healthy Skepticism: Assessing Realistic Model Performance. *Drug Discov Today*. 2009; 14(7):420 – 427. doi: <http://dx.doi.org/10.1016/j.drudis.2009.01.012>.
- [56] Abel R, Mondal S, Masse C, Greenwood J, Harriman G, Ashwell MA, Bhat S, Wester R, Frye L, Kapeller R, Friesner RA. Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. *Curr Opin Struct Biol*. 2017 Apr; 43(Supplement C):38–44.
- [57] Lovering F, Aevazelis C, Chang J, Dehnhardt C, Fitz L, Han S, Janz K, Lee J, Kaila N, McDonald J, Moore W, Moretto A, Papaioannou N, Richard D, Ryan MS, Wan ZK, Thorarensen A. Imidazotriazines: Spleen Tyrosine Kinase (Syk) Inhibitors Identified by Free-Energy Perturbation (FEP). *ChemMedChem*. 2016 Jan; 11(2):217–233.
- [58] Ciordia M, Pérez-Benito L, Delgado F, Trabanco AA, Tresadern G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. *Journal of chemical information and modeling*. 2016 Sep; 56(9):1856–1871.
- [59] Lenselink EB, Louvel J, Forti AF, van Veldhoven JPD, de Vries H, Mulder-Krieger T, McRobb FM, Negri A, Goose J, Abel R, van Vlijmen HWT, Wang L, Harder E, Sherman W, IJzerman AP, Beuming T. Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. *ACS omega*. 2016 Aug; 1(2):293–304.
- [60] Jorgensen WL. Computer-aided discovery of anti-HIV agents. *Bioorganic & medicinal chemistry*. 2016 Oct; 24(20):4768–4778.
- [61] Lin YL, Meng Y, Jiang W, Roux B. Explaining why Gleevec is a specific and potent inhibitor of Abl kinase. *Proc Natl Acad Sci U S A*. 2013 Jan; 110(5):1664–1669.
- [62] Lin YL, Meng Y, Huang L, Roux B. Computational Study of Gleevec and G6G Reveals Molecular Determinants of Kinase Inhibitor Selectivity. *J Am Chem Soc*. 2014 Oct; 136(42):14753–14762.
- [63] Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *J Am Chem Soc*. 2017 Jan; 139(2):946–957.
- [64] Shan Y, Eastwood MP, Zhang X, Kim ET, Arkhipov A, Dror RO, Jumper J, Kuriyan J, Shaw DE. Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell*. 2012 May; 149(4):860–870.
- [65] Sutto L, Gervasio FL. 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*. 2013 Jun; 110(26):10616–10621.
- [66] Park J, McDonald JJ, Petter RC, Houk KN. Molecular Dynamics Analysis of Binding of Kinase Inhibitors to WT EGFR and the T790M Mutant. *Journal of Chemical Theory and Computation*. 2016 Apr; 12(4):2066–2078.
- [67] Demir Ö, Baronio R, Salehi F, Wassman CD, Hall L, Hatfield GW, Chamberlin R, Kaiser P, Lathrop RH, Amaro RE. Ensemble-based computational approach discriminates functional activity of p53 cancer and rescue mutants. *PLoS Computational Biology*. 2011 Oct; 7(10):e1002238.

- [68] Witham S, Takano K, Schwartz C, Alexov E. A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. *Proteins*. 2011 Aug; 79(8):2444–2454.
- [69] Tsukamoto H, Farrens DL. A constitutively activating mutation alters the dynamics and energetics of a key conformational change in a ligand-free G protein-coupled receptor. *The Journal of biological chemistry*. 2013 Sep; 288(39):28207–28216.
- [70] Seeliger D, de Groot BL. Protein thermostability calculations using alchemical free energy simulations. *Biophysical journal*. 2010 May; 98(10):2309–2316.
- [71] Steinbrecher T, Zhu C, Wang L, Abel R, Negron C, Pearlman D, Feyfant E, Duan J, Sherman W. Predicting the Effect of Amino Acid Single-Point Mutations on Protein Stability—Large-Scale Validation of MD-Based Relative Free Energy Calculations. *Journal of Molecular Biology*. 2017 Apr; 429(7):948–963.
- [72] Ford MC, Babaoglu K. Examining the Feasibility of Using Free Energy Perturbation (FEP+) in Predicting Protein Stability. *Journal of chemical information and modeling*. 2017 Jun; 57(6):1276–1285.
- [73] Mondal J, Tiwary P, Berne BJ. How a Kinase Inhibitor Withstands Gatekeeper Residue Mutations. *Journal of the American Chemical Society*. 2016 Apr; 138(13):4608–4615.
- [74] Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J Am Chem Soc*. 2015 Feb; 137(7):2695–2703. doi: 10.1021/ja512751q.
- [75] Abel R, Wang L, Harder ED, Berne BJ, Friesner RA. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Accounts of chemical research*. 2017 Jul; 50(7):1625–1632.
- [76] Fan QW, Cheng CK, Nicolaides TP, Hackett CS, Knight ZA, Shokat KM, Weiss WA. A dual phosphoinositide-3-kinase alpha/mTOR inhibitor cooperates with blockade of epidermal growth factor receptor in PTEN-mutant glioma. *Cancer Res*. 2007 Sep; 67(17):7960–7965.
- [77] Kijima T, Shimizu T, Nonen S, Furukawa M, Otani Y, Minami T, Takahashi R, Hirata H, Nagatomo I, Takeda Y, Kida H, Goya S, Fujio Y, Azuma J, Tachibana I, Kawase I. Safe and successful treatment with erlotinib after gefitinib-induced hepatotoxicity: difference in metabolism as a possible mechanism. *J Clin Oncol*. 2011 Jul; 29(19):e588–90.
- [78] Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, Singh B, Heelan R, Rusch V, Fulton L, Mardis E, Kupfer D, Wilson R, Kris M, Varmus H. EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences*. 2004 Sep; 101(36):13306–13311.
- [79] Kim Y, Li Z, Apetri M, Luo B, Settleman JE, Anderson KS. Temporal resolution of autophosphorylation for normal and oncogenic forms of EGFR and differential effects of gefitinib. *Biochemistry*. 2012 Jun; 51(25):5212–5222.

- [80] Juchum M, Günther M, Laufer SA. Fighting Cancer Drug Resistance: Opportunities and Challenges for Mutation-Specific EGFR Inhibitors. *Drug Resist Updat.* 2015 May; 20:12–28. doi: [10.1016/j.drup.2015.05.002](https://doi.org/10.1016/j.drup.2015.05.002).
- [81] Din OS, Woll PJ. Treatment of gastrointestinal stromal tumor: focus on imatinib mesylate. *Ther Clin Risk Manag.* 2008 Feb; 4(1):149–162.
- [82] Lin YL, Roux B. Computational Analysis of the Binding Specificity of Gleevec to Abl, c-Kit, Lck, and c-Src Tyrosine Kinases. *J Am Chem Soc.* 2013 Oct; 135(39):14741–14753.
- [83] Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al-Lazikani B, Hersey A, Oprea TI, Overington JP. A Comprehensive Map of Molecular Drug Targets. *Nat Rev Drug Discov.* 2016 Dec; 16(1):19–34. doi: [10.1038/nrd.2016.230](https://doi.org/10.1038/nrd.2016.230).
- [84] Volkamer A, Eid S, Turk S, Jaeger S, Rippmann F, Fulle S. Pocketome of human kinases: prioritizing the ATP binding sites of (yet) untapped protein kinases for drug discovery. *J Chem Inf Model.* 2015 Mar; 55(3):538–549.
- [85] Wu P, Nielsen TE, Clausen MH. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol Sci.* 2015 Jul; 36(7):422–439.
- [86] Volkamer A, Eid S, Turk S, Rippmann F, Fulle S. Identification and Visualization of Kinase-Specific Subpockets. *J Chem Inf Model.* 2016 Feb; 56(2):335–346.
- [87] Christmann-Franck S, van Westen GJP, Papadatos G, Beltran Escudie F, Roberts A, Overington JP, Domine D. Unprecedentedly Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? *Journal of chemical information and modeling.* 2016 Sep; 56(9):1654–1675.
- [88] Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat Biotechnol.* 2011 Nov; 29(11):1039–1045.
- [89] Klaeger S, Heinzelmeir S, Wilhelm M, Polzer H, Vick B, Koenig PA, Reinecke M, Ruprecht B, Petzoldt S, Meng C, Zecha J, Reiter K, Qiao H, Helm D, Koch H, Schoof M, Canevari G, Casale E, Depaolini SR, Feuchtinger A, et al. The target landscape of clinical kinase drugs. *Science.* 2017 Dec; 358(6367).
- [90] Sun C, Hobor S, Bertotti A, Zecchin D, Huang S, Galimi F, Cottino F, Prahallas A, Grernrum W, Tzani A, Schlicker A, Wessels LFA, Smit EF, Thunnissen E, Halonen P, Lieftink C, Beijersbergen RL, Di Nicolantonio F, Bardelli A, Trusolino L, et al. Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction of ERBB3. *Cell Reports.* 2014 Apr; 7(1):86–93.
- [91] Manchado E, Weissmueller S, Morris JP, Chen CC, Wullenkord R, Lujambio A, de Stanchina E, Poirier JT, Gainor JF, Corcoran RB, Engelman JA, Rudin CM, Rosen N, Lowe SW. A combinatorial strategy for treating KRAS-mutant lung cancer. *Nature.* 2016 Jun; 534(7609):647–651.
- [92] Shao H, Shi S, Huang S, Hole AJ, Abbas AY, Baumli S, Liu X, Lam F, Foley DW, Fischer PM, Noble M, Endicott JA, Pepper C, Wang S. Substituted 4-(Thiazol-5-yl)-2-(phenylamino)pyrimidines Are Highly Active CDK9 Inhibitors: Synthesis, X-ray Crystal Structures, Structure–Activity Relationship, and Anticancer Activities. *J Med Chem.* 2013 Feb; 56(3):640–659.

- [93] Blake JF, Burkard M, Chan J, Chen H, Chou KJ, Diaz D, Dudley DA, Gaudino JJ, Gould SE, Grina J, Hunsaker T, Liu L, Martinson M, Moreno D, Mueller L, Orr C, Pacheco P, Qin A, Rasor K, Ren L, et al. Discovery of (S)-1-(1-(4-Chloro-3-fluorophenyl)-2-hydroxyethyl)-4-(2-((1-methyl-1H-pyrazol-5-yl)amino)pyrimidin-4-yl)pyridin-2(1H)-one (GDC-0994), an Extracellular Signal-Regulated Kinase 1/2 (ERK1/2) Inhibitor in Early Clinical Development. *J Med Chem.* 2016 Jun; 59(12):5650–5660.
- [94] Hole AJ, Baumli S, Shao H, Shi S, Huang S, Pepper C, Fischer PM, Wang S, Endicott JA, Noble ME. Comparative Structural and Functional Studies of 4-(Thiazol-5-yl)-2-(phenylamino)pyrimidine-5-carbonitrile CDK9 Inhibitors Suggest the Basis for Isootype Selectivity. *J Med Chem.* 2013 Feb; 56(3):660–670.
- [95] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr.* 2002 Jun; 58(Pt 6):899–907.
- [96] Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des.* 2013 Mar; 27(3):221–234.
- [97] Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, Kaus JW, Cerutti DS, Krilov G, Jorgensen WL, Abel R, Friesner RA. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J Chem Theory Comput.* 2016 Jan; 12(1):281–296.
- [98] Bochevarov AD, Harder E, Hughes TF, Greenwood JR, Braden DA, Philipp DM, Rinaldo D, Halls MD, Zhang J, Friesner RA. Jaguar: a high-performance quantum chemistry software program with strengths in life and materials sciences. *Int J Quantum Chem.* 2013; 113(18):2110–2142.
- [99] Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science.* 2016; 2:e55.
- [100] Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A, Belopolsky A, Bengio Y, Bergeron A, Bergstra J, Bisson V, Bleeker Snyder J, Bouchard N, Boulanger-Lewandowski N, Bouthillier X, de Brébisson A, Breuleux O, et al. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints.* 2016 May; abs/1605.02688. <http://arxiv.org/abs/1605.02688>.
- [101] Bosc N, Meyer C, Bonnet P. The use of novel selectivity metrics in kinase research. *BMC bioinformatics.* 2017 Jan; 18(1):17.
- [102] Cheng AC, Eksterowicz J, Geuns-Meyer S, Sun Y. Analysis of kinase inhibitor selectivity using a thermodynamics-based partition index. *J Med Chem.* 2010 Jun; 53(11):4502–4510.
- [103] Hauser K, Negron C, Albanese SK, Ray S, Steinbrecher T, Abel R, Chodera JD, Wang L. Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Communications Biology.* 2018 Jun; 1(1):70.
- [104] Hari SB, Merritt EA, Maly DJ. Sequence determinants of a specific inactive protein kinase conformation. *Chemistry & biology.* 2013 Jun; 20(6):806–815.

- [105] Hu J, Ahuja LG, Meharena HS, Kannan N, Kornev AP, Taylor SS, Shaw AS. Kinase regulation by hydrophobic spine assembly in cancer. *Molecular and cellular biology*. 2015 Jan; 35(1):264–276.
- [106] Laplante M, Sabatini DM. mTOR signaling in growth control and disease. *Cell*. 2012 Apr; 149(2):274–293.
- [107] Saxton RA, Sabatini DM. mTOR Signaling in Growth, Metabolism, and Disease. *Cell*. 2017 Mar; 168(6):960–976.
- [108] Kim DH, Sarbassov DD, Ali SM, King JE, Latek RR, Erdjument-Bromage H, Tempst P, Sabatini DM. mTOR interacts with raptor to form a nutrient-sensitive complex that signals to the cell growth machinery. *Cell*. 2002 Jul; 110(2):163–175.
- [109] Hara K, Maruki Y, Long X, Yoshino Ki, Oshiro N, Hidayat S, Tokunaga C, Avruch J, Yonezawa K. Raptor, a binding partner of target of rapamycin (TOR), mediates TOR action. *Cell*. 2002 Jul; 110(2):177–189.
- [110] Yang H, Jiang X, Li B, Yang HJ, Miller M, Yang A, Dhar A, Pavletich NP. Mechanisms of mTORC1 activation by RHEB and inhibition by PRAS40. *Nature*. 2017 Dec; 552(7685):368–373.
- [111] Sarbassov DD, Ali SM, Kim DH, Guertin DA, Latek RR, Erdjument-Bromage H, Tempst P, Sabatini DM. Rictor, a novel binding partner of mTOR, defines a rapamycin-insensitive and raptor-independent pathway that regulates the cytoskeleton. *Current biology : CB*. 2004 Jul; 14(14):1296–1302.
- [112] Bar-Peled L, Sabatini DM. SnapShot: mTORC1 signaling at the lysosomal surface. *Cell*. 2012 Dec; 151(6):1390–1390.e1.
- [113] Peterson TR, Laplante M, Thoreen CC, Sancak Y, Kang SA, Kuehl WM, Gray NS, Sabatini DM. DEPTOR is an mTOR inhibitor frequently overexpressed in multiple myeloma cells and required for their survival. *Cell*. 2009 May; 137(5):873–886.
- [114] Yang H, Rudge DG, Koos JD, Vaidalingam B, Yang HJ, Pavletich NP. mTOR kinase structure, mechanism and regulation. *Nature*. 2013 May; 497(7448):217–223.
- [115] Gaubitz C, Oliveira TM, Prouteau M, Leitner A, Karuppasamy M, Konstantinidou G, Rispal D, Eltschinger S, Robinson GC, Thore S, Aebersold R, Schaffitzel C, Loewith R. Molecular Basis of the Rapamycin Insensitivity of Target Of Rapamycin Complex 2. *Molecular cell*. 2015 Jun; 58(6):977–988.
- [116] Aylett CHS, Sauer E, Imseng S, Boehringer D, Hall MN, Ban N, Maier T. Architecture of human mTOR complex 1. *Science*. 2016 Jan; 351(6268):48–52.
- [117] Inoki K, Li Y, Zhu T, Wu J, Guan KL. TSC2 is phosphorylated and inhibited by Akt and suppresses mTOR signalling. *Nature cell biology*. 2002 Sep; 4(9):648–657.
- [118] Manning BD, Tee AR, Logsdon MN, Blenis J, Cantley LC. Identification of the tuberous sclerosis complex-2 tumor suppressor gene product tuberin as a target of the phosphoinositide 3-kinase/akt pathway. *Molecular cell*. 2002 Jul; 10(1):151–162.
- [119] Jones RG, Plas DR, Kubek S, Buzzai M, Mu J, Xu Y, Birnbaum MJ, Thompson CB. AMP-activated protein kinase induces a p53-dependent metabolic checkpoint. *Molecular cell*. 2005 Apr; 18(3):283–293.

- [120] Inoki K, Li Y, Xu T, Guan KL. Rheb GTPase is a direct target of TSC2 GAP activity and regulates mTOR signaling. *Genes & development*. 2003 Aug; 17(15):1829–1834.
- [121] Bar-Peled L, Schweitzer LD, Zoncu R, Sabatini DM. Ragulator is a GEF for the rag GTPases that signal amino acid levels to mTORC1. *Cell*. 2012 Sep; 150(6):1196–1208.
- [122] Kim E, Goraksha-Hicks P, Li L, Neufeld TP, Guan KL. Regulation of TORC1 by Rag GTPases in nutrient response. *Nature cell biology*. 2008 Aug; 10(8):935–945.
- [123] Sancak Y, Bar-Peled L, Zoncu R, Markhard AL, Nada S, Sabatini DM. Ragulator-Rag complex targets mTORC1 to the lysosomal surface and is necessary for its activation by amino acids. *Cell*. 2010 Apr; 141(2):290–303.
- [124] Efeyan A, Zoncu R, Sabatini DM. Amino acids and mTORC1: from lysosomes to disease. *Trends in molecular medicine*. 2012 Sep; 18(9):524–533.
- [125] Hay N, Sonenberg N. Upstream and downstream of mTOR. *Genes & development*. 2004 Aug; 18(16):1926–1945.
- [126] Lamming DW, Sabatini DM. A Central role for mTOR in lipid homeostasis. *Cell metabolism*. 2013 Oct; 18(4):465–469.
- [127] Delarue M, Brittingham GP, Pfeffer S, Surovtsev IV, Pinglay S, Kennedy KJ, Schaffer M, Gutierrez JI, Sang D, Poterewicz G, Chung JK, Plitzko JM, Groves JT, Jacobs-Wagner C, Engel BD, Holt LJ. mTORC1 Controls Phase Separation and the Biophysical Properties of the Cytoplasm by Tuning Crowding. *Cell*. 2018 Jun; .
- [128] Ballou LM, Lin RZ. Rapamycin and mTOR kinase inhibitors. *Journal of chemical biology*. 2008 Nov; 1(1-4):27–36.
- [129] Lamming DW, Ye L, Sabatini DM, Baur JA. Rapalogs and mTOR inhibitors as anti-aging therapeutics. *The Journal of clinical investigation*. 2013 Mar; 123(3):980–989.
- [130] Hudes G, Carducci M, Tomczak P, Dutcher J, Figlin R, Kapoor A, Staroslawska E, Sosman J, McDermott D, Bodrogi I, Kovacevic Z, Lesovoy V, Schmidt-Wolf IGH, Barbarash O, Gokmen E, O'Toole T, Lustgarten S, Moore L, Motzer RJ, Global ARCC Trial. Temsirolimus, interferon alfa, or both for advanced renal-cell carcinoma. *The New England journal of medicine*. 2007 May; 356(22):2271–2281.
- [131] Motzer RJ, Escudier B, Oudard S, Hutson TE, Porta C, Bracarda S, Grünwald V, Thompson JA, Figlin RA, Hollaender N, Urbanowitz G, Berg WJ, Kay A, Lebwohl D, Ravaud A, RECORD-1 Study Group. Efficacy of everolimus in advanced renal cell carcinoma: a double-blind, randomised, placebo-controlled phase III trial. *Lancet (London, England)*. 2008 Aug; 372(9637):449–456.
- [132] Hausch F, Kozany C, Theodoropoulou M, Fabian AK. FKBP_s and the Akt/mTOR pathway. *Cell cycle (Georgetown, Tex)*. 2013 Aug; 12(15):2366–2370.
- [133] Yip CK, Murata K, Walz T, Sabatini DM, Kang SA. Structure of the human mTOR complex I and its implications for rapamycin inhibition. *Molecular cell*. 2010 Jun; 38(5):768–774.
- [134] Schreiber KH, Ortiz D, Academia EC, Anies AC, Liao CY, Kennedy BK. Rapamycin-mediated mTORC2 inhibition is determined by the relative expression of FK506-binding proteins. *Aging cell*. 2015 Apr; 14(2):265–273.

- [135] Slotkin EK, Patwardhan PP, Vasudeva SD, de Stanchina E, Tap WD, Schwartz GK. MLN0128, an ATP-competitive mTOR kinase inhibitor with potent in vitro and in vivo antitumor activity, as potential therapy for bone and soft-tissue sarcoma. *Molecular cancer therapeutics*. 2015 Feb; 14(2):395–406.
- [136] Hassan B, Akcakanat A, Sangai T, Evans KW, Adkins F, Eterovic AK, Zhao H, Chen K, Chen H, Do KA, Xie SM, Holder AM, Naing A, Mills GB, Meric-Bernstam F. Catalytic mTOR inhibitors can overcome intrinsic and acquired resistance to allosteric mTOR inhibitors. *Oncotarget*. 2014 Sep; 5(18):8544–8557.
- [137] García-García C, Ibrahim YH, Serra V, Calvo MT, Guzmán M, Grueso J, Aura C, Pérez J, Jessen K, Liu Y, Rommel C, Tabernero J, Baselga J, Scaltriti M. Dual mTORC1/2 and HER2 blockade results in antitumor activity in preclinical models of breast cancer resistant to anti-HER2 therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2012 May; 18(9):2603–2612.
- [138] Chresta CM, Davies BR, Hickson I, Harding T, Cosulich S, Critchlow SE, Vincent JP, Ellston R, Jones D, Sini P, James D, Howard Z, Dudley P, Hughes G, Smith L, Maguire S, Hummersone M, Malagu K, Menear K, Jenkins R, et al. AZD8055 is a potent, selective, and orally bioavailable ATP-competitive mammalian target of rapamycin kinase inhibitor with in vitro and in vivo antitumor activity. *Cancer research*. 2010 Jan; 70(1):288–298.
- [139] Mallon R, Feldberg LR, Lucas J, Chaudhary I, Dehnhardt C, Santos ED, Chen Z, dos Santos O, Ayral-Kaloustian S, Venkatesan A, Hollander I. Antitumor efficacy of PKI-587, a highly potent dual PI3K/mTOR kinase inhibitor. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2011 May; 17(10):3193–3203.
- [140] Mukherjee B, Tomimatsu N, Amacherla K, Camacho CV, Pichamoorthy N, Burma S. The Dual PI3K/mTOR Inhibitor NVP-BEZ235 Is a Potent Inhibitor of ATM- and DNA-PKCs-Mediated DNA Damage Responses. *Neoplasia*. 2012 Jan; 14(1):34–IN8.
- [141] Pongas G, Fojo T. BEZ235: When Promising Science Meets Clinical Reality. *The oncologist*. 2016 Sep; 21(9):1033–1034.
- [142] Eichhorn PJA, Gili M, Scaltriti M, Serra V, Guzman M, Nijkamp W, Beijersbergen RL, Valero V, Seoane J, Bernards R, Baselga J. Phosphatidylinositol 3-kinase hyperactivation results in lapatinib resistance that is reversed by the mTOR/phosphatidylinositol 3-kinase inhibitor NVP-BEZ235. *Cancer research*. 2008 Nov; 68(22):9221–9230.
- [143] Rodrik-Outmezguine VS, Chandarlapaty S, Pagano NC, Poulikakos PI, Scaltriti M, Moskatek E, Baselga J, Guichard S, Rosen N. mTOR kinase inhibition causes feedback-dependent biphasic regulation of AKT signaling. *Cancer discovery*. 2011 Aug; 1(3):248–259.
- [144] Fan QW, Knight ZA, Goldenberg DD, Yu W, Mostov KE, Stokoe D, Shokat KM, Weiss WA. A dual PI3 kinase/mTOR inhibitor reveals emergent efficacy in glioma. *Cancer Cell*. 2006 May; 9(5):341–349.
- [145] Wagle N, Grabiner BC, Van Allen EM, Hodis E, Jacobus S, Supko JG, Stewart M, Choueiri TK, Gandhi L, Cleary JM, Elfiky AA, Taplin ME, Stack EC, Signoretti S, Loda M, Shapiro GI, Sabatini DM, Lander ES, Gabriel SB, Kantoff PW, et al. Activating mTOR mutations in a patient with an extraordinary response on a phase I trial of everolimus and pazopanib. *Cancer discovery*. 2014 May; 4(5):546–553.

- [146] Wagle N, Nikhil, Grabiner BC, Van Allen E, Amin-Mansour A, Taylor-Weiner A, Amaro R, Rosenberg M, Mara G, Nathanael B, Justine A, Guo Y, Yanan S, Swanson S, J. Ruan D, T. Hanna G, Glenn J, Haddad R, Robert I, Getz G, Gad K, Kwiatkowski D, David J, Carter S, L. Sabatini D, David M, Jänne P, A, Garraway L, Levi A, Lorch J, Jochen H. Response and acquired resistance to everolimus in anaplastic thyroid cancer. *The New England journal of medicine*. 2014 Oct; 371(15):1426–1433.
- [147] Rodrik-Outmezguine VS, Okaniwa M, Yao Z, Novotny CJ, McWhirter C, Banaji A, Won H, Wong W, Berger M, de Stanchina E, Barratt DG, Cosulich S, Klinowska T, Rosen N, Shokat KM. Overcoming mTOR resistance mutations with a new-generation mTOR inhibitor. *Nature*. 2016 Jun; 534(7606):272–276.
- [148] Fan Q, Aksoy O, Wong RA, Ilkhanizadeh S, Novotny CJ, Gustafson WC, Truong AYQ, Cayanan G, Simonds EF, Haas-Kogan D, Phillips JJ, Nicolaides T, Okaniwa M, Shokat KM, Weiss WA. A Kinase Inhibitor Targeted to mTORC1 Drives Regression in Glioblastoma. *Cancer Cell*. 2017 Mar; 31(3):424–435.
- [149] Guertin DA, Sabatini DM. Defining the role of mTOR in cancer. *Cancer Cell*. 2007 Jul; 12(1):9–22.
- [150] Grabiner BC, Nardi V, Birsoy K, Possemato R, Shen K, Sinha S, Jordan A, Beck AH, Sabatini DM. A diverse array of cancer-associated MTOR mutations are hyperactivating and can predict rapamycin sensitivity. *Cancer discovery*. 2014 May; 4(5):554–563.
- [151] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*. 2013 Jul; 499(7456):43–49.
- [152] Xu J, Pham CG, Albanese SK, Dong Y, Oyama T, Lee CH, Rodrik-Outmezguine V, Yao Z, Han S, Chen D, Parton DL, Chodera JD, Rosen N, Cheng EH, Hsieh JJ. Mechanistically distinct cancer-associated mTOR activation clusters predict sensitivity to rapamycin. *The Journal of clinical investigation*. 2016 Sep; 126(9):3526–3540.
- [153] Salsbury FR. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current opinion in pharmacology*. 2010 Dec; 10(6):738–744.
- [154] Shirts M, Pande VS. COMPUTING: Screen Savers of the World Unite! *Science*. 2000 Dec; 290(5498):1903–1904.
- [155] Stehr H, Jang SHJ, Duarte JM, Wierling C, Lehrach H, Lappe M, Lange BMH. The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Molecular cancer*. 2011 May; 10(1):54.
- [156] Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical Free Energy Methods for Drug Discovery: Progress and Challenges. *Curr Opin Struct Biol*. 2011 Apr; 21(2):150–160. doi: [10.1016/j.sbi.2011.01.011](https://doi.org/10.1016/j.sbi.2011.01.011).
- [157] Parton DL, Grinaway PB, Hanson SM, Beauchamp KA, Chodera JD. Ensembler: Enabling High-Throughput Molecular Simulations at the Superfamily Scale. *PLoS computational biology*. 2016 Jun; 12(6):e1004728.
- [158] Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*. 2010 Jun; 78(8):1950–1958.

- [159] Jorgensen, William L, Chandrasekhar, Jayaraman, Madura, Jeffry D, Impey, Roger W, Klein, Michael L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*. 1998 Aug; 79(2):926–935.
- [160] Eastman, Peter, Swails, Jason, Chodera, John D, McGibbon, Robert T, Zhao, Yutong, Beauchamp, Kyle A, Wang, Lee-Ping, Simmonett, Andrew C, Harrigan, Matthew P, Stern, Chaya D, Wiewiora, Rafal P, Brooks, Bernard R, Pande, Vijay S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*. 2017 Jul; 13(7):e1005659.
- [161] Eastman P, Pande VS. CCMA: A Robust, Parallelizable Constraint Method for Molecular Simulations. *Journal of chemical theory and computation*. 2010 Feb; 6(2):434–437.
- [162] Miyamoto S, Kollman PA. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*. 1992; 13(8):952–962.
- [163] Eastman P, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, Beauchamp KA, Lane TJ, Wang LP, Shukla D, Tye T, Houston M, Stich T, Klein C, Shirts MR, Pande VS. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation*. 2013 Jan; 9(1):461–469.
- [164] McGibbon, Robert T, Beauchamp, Kyle A, Harrigan, Matthew P, Klein, Christoph, Swails, Jason M, Hernández, Carlos X, Schwantes, Christian R, Wang, Lee-Ping, Lane, Thomas J, Pande, Vijay S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical journal*. 2015 Oct; 109(8):1528–1532.
- [165] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry*. 2004 Mar; 47(7):1739–1749.
- [166] Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry*. 2004 Mar; 47(7):1750–1759.
- [167] Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of medicinal chemistry*. 2006 Oct; 49(21):6177–6196.
- [168] Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J Am Chem Soc*. 2015 Feb; 137(7):2695–2703.
- [169] Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J. Interaction Models for Water in Relation to Protein Hydration. In: *Intermolecular Forces* Dordrecht: Springer, Dordrecht; 1981.p. 331–342.

- [170] Bennett CH. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*. 1976 Oct; 22(2):245–268.
- [171] Shirts MR, Bair E, Hooker G, Pande VS. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Physical Review Letters*. 2003; 91(14).
- [172] Shukla D, Meng Y, Roux B, Pande VS. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature communications*. 2014; 5:3397.
- [173] Jura N, Zhang X, Endres NF, Seeliger MA, Schindler T, Kuriyan J. Catalytic Control in the EGF Receptor and Its Connection to General Kinase Regulatory Mechanisms. *Molecular cell*. 2011 Apr; 42(1):9–22.
- [174] Urano J, Sato T, Matsuo T, Otsubo Y, Yamamoto M, Tamanoi F. Point mutations in TOR confer Rheb-independent growth in fission yeast and nutrient-independent mammalian TOR signaling in mammalian cells. *Proceedings of the National Academy of Sciences*. 2007 Feb; 104(9):3514–3519.
- [175] Clark AJ, Gindin T, Zhang B, Wang L, Abel R, Murret CS, Xu F, Bao A, Lu NJ, Zhou T, et al. Free Energy Perturbation Calculation of Relative Binding Free Energy between Broadly Neutralizing Antibodies and the gp120 Glycoprotein of HIV-1. *Journal of molecular biology*. 2017; 429(7):930–947.
- [176] Steinbrecher T, Zhu C, Wang L, Abel R, Negron C, Pearlman D, Feyfant E, Duan J, Sherman W. Predicting the Effect of Amino Acid Single-Point Mutations on Protein Stability—Large-Scale Validation of MD-Based Relative Free Energy Calculations. *Journal of molecular biology*. 2017; 429(7):948–963.
- [177] Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM, Hellmann MD, Barron DA, Schram AM, Hameed M, Dogan S, Ross DS, Hechtman JF, DeLair DF, Yao J, Mandelker DL, et al. Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients. *Nat Med*. 2017 May; 23(6):703–713. doi: [10.1038/nm.4333](https://doi.org/10.1038/nm.4333).
- [178] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*. 2012 May; 2(5):401–404.
- [179] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science signaling*. 2013 Apr; 6(269):pl1–pl1.
- [180] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012 Mar; 483(7391):603–607.
- [181] Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, BC Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, Case Western Reserve University, Dana-Farber Cancer Institute, Duke University, Greater Poland Cancer Centre, Harvard Medical School, Institute

- for Systems Biology, KU Leuven, Mayo Clinic, Memorial Sloan Kettering Cancer Center, National Cancer Institute, Nationwide Children's Hospital, Stanford University, University of Alabama, University of Michigan, et al. Integrated genomic characterization of oesophageal carcinoma. *Nature*. 2017 Jan; 541(7636):169–175.
- [182] Stuttfeld E, Aylett CH, Imseng S, Boehringer D, Scaiola A, Sauer E, Hall MN, Maier T, Ban N. Architecture of the human mTORC2 core complex. *eLife*. 2018 Feb; 7:48.
- [183] Yuan HX, Guan KL. Structural insights of mTOR complex 1. *Cell research*. 2016 Mar; 26(3):267–268.
- [184] Yang H, Wang J, Liu M, Chen X, Huang M, Tan D, Dong MQ, Wong CCL, Wang J, Xu Y, Wang HW. 4.4 Resolution Cryo-EM structure of human mTOR Complex 1. *Protein & cell*. 2016 Dec; 7(12):878–887.
- [185] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science*. 2002; 298:1912–1934.
- [186] Hunter T, Manning G. The Eukaryotic Protein Kinase Superfamily and the Emergence of Receptor Tyrosine Kinases. In: Wheeler DL, Yarden Y, editors. *Receptor Tyrosine Kinases: Structure, Functions and Role in Human Disease* New York, NY: Springer New York; 2015.p. 1–15.
- [187] Knight ZA, Lin H, Shokat KM. Targeting the cancer kinase through polypharmacology. *Nature Reviews Cancer*. 2010 Feb; 10(2):130–137. <http://www.nature.com.proxy.library.cornell.edu/nrc/journal/v10/n2/full/nrc2787.html>, doi: 10.1038/nrc2787.
- [188] American Cancer Society, Cancer Facts & Figures 2015. American Cancer Society; 2015.
- [189] Cohen P, Tcherpakov M. Will the Ubiquitin System Furnish as Many Drug Targets as Protein Kinases? *Cell*. 2010 Nov; 143(5):686–693. <http://linkinghub.elsevier.com/retrieve/pii/S0092867410012961>, doi: 10.1016/j.cell.2010.11.016.
- [190] Stegmeier F, Warmuth M, Sellers WR, Dorsch M. Targeted cancer therapies in the twenty-first century: Lessons from imatinib. *Clin Pharm & Therap*. 2010; 87:543–552. doi: 10.1038/clpt.2009.297.
- [191] Wu P, Nielsen TE, Clausen MH. FDA-approved small-molecule kinase inhibitors. *Trends in Pharmacological Sciences*. 2015 Jul; 36(7):422–439. <http://linkinghub.elsevier.com/retrieve/pii/S0165614715000772>, doi: 10.1016/j.tips.2015.04.005.
- [192] Volkamer A, Eid S, Turk S, Jaeger S, Rippmann F, Fulle S. Pocketome of Human Kinases: Prioritizing the ATP Binding Sites of (Yet) Untapped Protein Kinases for Drug Discovery. *Journal of Chemical Information and Modeling*. 2015; 55(3):538–549. <http://dx.doi.org/10.1021/ci500624s>, doi: 10.1021/ci500624s, pMID: 25557645.
- [193] Chandarlapaty S, Sawai A, Scaltriti M, Rodrik-Outmezguine V, Grbovic-Huezo O, Serra V, Majumder PK, Baselga J, Rosen N. AKT Inhibition Relieves Feedback Suppression of Receptor Tyrosine Kinase Expression and Activity. *Cancer Cell*. 2011 Jan; 19(1):58–71. <http://linkinghub.elsevier.com/retrieve/pii/S1535610810004332>, doi: 10.1016/j.ccr.2010.10.031.

- [194] Karaman MW, Herrgard S, Treiber DK, Gallant P, Atteridge CE, Campbell BT, Chan KW, Ciceri P, Davis MI, Edeen PT, Faraoni R, Floyd M, Hunt JP, Lockhart DJ, Milanov ZV, Morrison MJ, Pallares G, Patel HK, Pritchard S, Wodicka LM, et al. A quantitative analysis of kinase inhibitor selectivity. *Nature Biotech.* 2008; 26:127–132.
- [195] Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology.* 2011 Nov; 29(11):1046–1051.
- [196] Médard G, Pachl F, Ruprecht B, Klaeger S, Heinzelmeir S, Helm D, Qiao H, Ku X, Wilhelm M, Kuehne T, Wu Z, Dittmann A, Hopf C, Kramer K, Kuster B. Optimized chemical proteomics assay for kinase inhibitor profiling. *Journal of proteome research.* 2015 Mar; 14(3):1574–1586.
- [197] Chambers SP, Austen DA, Fulghum JR, Kim WM. High-throughput screening for soluble recombinant expressed kinases in *Escherichia coli* and insect cells. *Protein Expression and Purification.* 2004; 36:40–47. doi: [10.1016/j.pep.2004.03.003](https://doi.org/10.1016/j.pep.2004.03.003).
- [198] Wang L, Foster M, Zhang Y, Tschantz WR, Yang L, Worrall J, Loh C, Xu X. High yield expression of non-phosphorylated protein tyrosine kinases in insect cells. *Protein Express Pur.* 2008; 61:204–211. doi: [10.1016/j.pep.2008.05.017](https://doi.org/10.1016/j.pep.2008.05.017).
- [199] Seeliger MA, Young M, Henderson MN, Pellicena P, King DS, Falick AM, Kuriyan J. High yield bacterial expression of active c-Abl and c-Src tyrosine kinases. *Protein Sci.* 2005; 14:3135–3139.
- [200] The Structural Genomics Consortium, SGC: Progress in protein kinase structural biology; 2012. <http://www.thescg.org/scientists/resources/kinases>, updated 18 Jul 2012.
- [201] Weisberg E, Manley PW, Breitenstein W, Brüggen J, Cowan-Jacob SW, Ray A, Huntly B, Fabbro D, Fendrich G, Hall-Meyers E, Kung AL, Mestan J, Daley GQ, Callahan L, Catley L, Cavazza C, Azam M, Mohammed A, Neuberg D, Wright RD, et al. Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer Cell.* 2005 Feb; 7(2):129–141.
- [202] Levinson NM, Kuchment O, Shen K, Young MA, Koldobskiy M, Karplus M, Cole PA, Kuriyan J. A Src-like inactive conformation in the Abl tyrosine kinase domain. *Plos Biology.* 2006; 4:e144.
- [203] Horio T, Hamasaki T, Inoue T, Wakayama T. Structural factors contributing to the Abl/Lyn dual inhibitory activity of 3-substituted benzamide derivatives. *Bioorganic & medicinal* 2007; .
- [204] UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucl Acids Res.* 2017; 45:D158–D169. doi: [10.1093/nar/gkh131](https://doi.org/10.1093/nar/gkh131).
- [205] Puhl AC, Giacomini C, Irazoqui G, Batista Viera F, Villarino A, Terenzi H. Covalent immobilization of tobacco-etch-virus NIa protease: a useful tool for cleavage of the histidine tag of recombinant proteins. *Biotechnology and Applied Biochemistry.* 2009 May; 53(3):165–174.
- [206] Nallamsetty S, Kapust RB, Tözsér J, Cherry S, Tropea JE, Copeland TD, Waugh DS. Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro. *Protein expression and purification.* 2004 Nov; 38(1):108–115.

- [207] Sun C, Liang J, Shi R, Gao X, Zhang R, Hong F, Yuan Q, Wang S. Tobacco etch virus protease retains its activity in various buffers and in the presence of diverse additives. *Protein expression and purification*. 2012 Mar; 82(1):226–231.
- [208] Lo MC, Aulabaugh A, Jin G, Cowling R, Bard J, Malamas M, Ellestad G. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Analytical Biochemistry*. 2004 Sep; 332(1):153–159.
- [209] Ericsson UB, Hallberg BM, Detitta GT, Dekker N, Nordlund P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical Biochemistry*. 2006 Oct; 357(2):289–298.
- [210] Matulis D, Kranz JK, Salemme FR, Todd MJ. Thermodynamic stability of carbonic anhydrase: measurements of binding affinity and stoichiometry using ThermoFluor. *Biochemistry*. 2005 Apr; 44(13):5258–5266.
- [211] Iyer GH, Taslimi P, Pazhanisamy S. Staurosporine-based binding assay for testing the affinity of compounds to protein kinases. *Analytical Biochemistry*. 2008 Feb; 373(2):197–206.
- [212] Levinson NM, Boxer SG. Structural and Spectroscopic Analysis of the Kinase Inhibitor Bosutinib and an Isomer of Bosutinib Binding to the Abl Tyrosine Kinase Domain. *PLoS One*. 2012; 7:e29828.
- [213] Levinson NM, Boxer SG. A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity. *Nature chemical biology*. 2014 Feb; 10(2):127–132.
- [214] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery*. 2012 May; 2(5):401–404. doi: [10.1158/2159-8290.CD-12-0095](https://doi.org/10.1158/2159-8290.CD-12-0095).
- [215] Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Brannon AR, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman JF, Yao J, Song W, Ross DS, Oultache A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *The Journal of Molecular Diagnostics*. 2015 May; 17(3):251–264. doi: [10.1016/j.jmoldx.2014.12.006](https://doi.org/10.1016/j.jmoldx.2014.12.006).
- [216] Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. . 2007; 8(11):R232. <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-11-r232>, doi: 10.1186/gb-2007-8-11-r232.
- [217] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*. 2011; 39(17):e118. <http://dx.doi.org/10.1093/nar/gkr407>, doi: 10.1093/nar/gkr407.
- [218] Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G. Oncotator: cancer variant annotation tool. *Human mutation*. 2015 Apr; 36(4):E2423–9.
- [219] Klock HE, Koesema EJ, Knuth MW, Lesley SA. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins: Structure, Function, and Bioinformatics*. 2008 May; 71(2):982–994. doi: [10.1002/prot.21786](https://doi.org/10.1002/prot.21786).

- [220] Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ, Kleywegt GJ. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research. 2013; 41(D1):D483. <http://dx.doi.org/10.1093/nar/gks1258>, doi: 10.1093/nar/gks1258.
- [221] Sørensen HP, Mortensen KK. Advanced genetic strategies for recombinant protein expression in Escherichia coli. Journal of Biotechnology. 2005; 115(2):113 – 128. <http://www.sciencedirect.com/science/article/pii/S0168165604004560>, doi: <http://dx.doi.org/10.1016/j.jbiotec.2004.08.004>.
- [222] Haacke A, Fendrich G, Ramage P, Geiser M. Chaperone Over-Expression in Escherichia Coli: Apparent Increased Yields of Soluble Recombinant Protein Kinases Are Due Mainly to Soluble Aggregates. Protein Expr Purif. 2009 Apr; 64(2):185–193. doi: [10.1016/j.pep.2008.10.022](https://doi.org/10.1016/j.pep.2008.10.022).
- [223] Kim DM, Swartz JR. Prolonging Cell-Free Protein Synthesis with a Novel ATP Regeneration System. Biotechnol Bioeng. 1999; 66(3):180–188. doi: [10.1002/\(SICI\)1097-0290\(1999\)66:3<180::AID-BIT6>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1097-0290(1999)66:3<180::AID-BIT6>3.0.CO;2-S).
- [224] Sawasaki T, Ogasawara T, Morishita R, Endo Y. A Cell-Free Protein Synthesis System for High-Throughput Proteomics. Proc Natl Acad Sci. 2002 Nov; 99(23):14652–14657. doi: [10.1073/pnas.232580399](https://doi.org/10.1073/pnas.232580399).