INTEGRATION OF LIGAND- AND STRUCTURE-BASED CHEMINFORMATICS TOOLS WITH PROTEIN DYNAMIC MODELING FOR DRUG DESIGN

By

Benjamin Patrick Brown

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

June 30, 2022

Nashville, Tennessee

Approved:

Ambra Pozzi, Ph.D.

Hassane Mchaourab, Ph.D.

Jens Meiler, Ph.D.

Jarrod Smith, Ph.D.

Charles Weaver, Ph.D.

To my parents, Terrence and Jill Brown,
for filling my life with opportunities and supporting me in all my endeavors,
and
to my beloved wife, Grace Hsu,
for being my best friend and life partner

ACKNOWLEDGMENTS

When I first rotated in the Meiler Lab in 2015, I could barely login to my assigned Linux workstation. I did not know anything about programming in any language, let alone C++. All I knew was that I wanted to learn how to do computational research in chemistry and structural biology. Despite my clear unpreparedness for research in the Meiler Lab, Jens accepted me into the group anyway. It is because Jens believed that I could be successful in his lab that I am writing this dissertation. So, first and foremost, I need to thank my doctoral advisor, Dr. Jens Meiler, PhD, for his constant and committed mentorship. Jens is brilliant – not just because of his intellectual power, but also because of his vision and the extent of his perspective. Jens is a true scholar, reveling not just in his own scientific achievements, but also in the academic and personal successes of his students. I am honored to be his pupil, and I will always be grateful that he took a chance on me.

I have also had the great fortune of being supported by an outstanding committee of scientists. I am very grateful to all of my committee members – Drs. Ambra Pozzi, Hassane Mchaourab, Jarrod Smith, and Charles Weaver. Throughout my PhD they have provided me with balanced guidance, perspective, and insight into my research that I otherwise would not have had. I am particularly indebted to my Committee Chair, Dr. Ambra Pozzi, PhD, for involving me in collaborations, directly supporting and co-mentoring some of my earliest graduate research, and for including me in training opportunities outside of research. I also would like to say thank you to Dr. Terry Lybrand, who formerly served on my committee. Though he has now retired, Dr. Lybrand provided me with many helpful conversations on different methods for estimating protein-ligand binding energies.

I need to acknowledge Dr. Christine Lovly, MD, PhD. Christine is the truest embodiment of a physician-scientist, and I have been fortunate to work with her and her team over the last few years in the precision oncology / personalized structural biology interface. More than just a collaborator, she has been an invaluable guide for my growth as a person.

I need to also thank Drs. Adam Smith, Soyeon Kim, David Westover, Zhenfang Du, and Yunkai Zhang. Everything I may have simulated in the computer, these outstanding people brought to life at the bench. They constantly challenged me to develop a deeper understanding of the biology we studied, and I am grateful for their comradery.

In addition to my faculty mentors and advisors, I need to say a very special thank you to Jeffrey Mendenhall. I would not have learned or accomplished as much as I have without his mentorship. Somebody had to physically teach me to use my Linux workstation. Somebody had to teach me to program. Somebody had to filter all of my half-baked ideas and help me identify the ones to continue working on. For my day-to-day activities, this person was Jeff. We spent many evenings working on code, tossing ideas back and forth, in our shared office. I will always be grateful for his seemingly limitless supply of patience.

My life did not begin to exist when I came to Vanderbilt, and I need to thank my parents for everything that they have done to get me here. It is impossible to say in just a few words all of the ways in which my mom and dad have supported, and continue to support, me in all that I do. This dissertation belongs as much to them as it does to me.

Finally, I need to thank my wife, Grace. For half of my life now, Grace has supported me as a friend, girlfriend, fiancée, wife, and/or soon-to-be parent. It is impossible for me to imagine my life without her because she has been with me for so much of it. In 2015, she moved to Philadelphia for work while I moved to Nashville. In 2018, we got married, and Grace moved to Nashville so that I could complete my studies without disruption. I am indebted to my wife for her sacrifice, and I am humbled by the magnitude of her love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Summary

Computer-aided drug design (CADD) has become a core component of modern drug discovery (Macalino et al., 2015). CADD is typically separated into two categories: ligand-based (LB) and structure-based (SB) (Sliwoski et al., 2014). LB methods do not require information on mechanism of action to yield predictions of molecular properties of interest (e.g., biological activity on a target receptor, solubility, etc.) and are frequently utilized in small molecule virtual high-throughput screening (vHTS). LB models are built using information from datasets of existing ligands, and therefore prediction quality is dependent on the quality and volume of training data. Quantitative structure-activity relationship (QSAR) modeling, which mathematically relates chemical descriptors of small molecules to properties of interest, has emerged as a powerful approach to leverage continual advancements in machine learning (ML) (Yang et al., 2019).

SB methods model interactions between ligands and target receptors. The interaction score predicts the activity of the ligand on the target. The primary benefits of SB methods are arguably twofold: (1) they do not require training data and thus in principle can be applied indiscriminately to any target receptor; (2) they are generally chemically intuitive and can guide rational design. There are also arguably two significant challenges associated with SB methods: (1) ranking compounds based on interaction scores first requires determination of the biologically relevant mode of interaction, the lack of which leads to substantial error in compound ranking; (2) they are orders of magnitude more computationally expensive than LB methods using even the simplest approaches, with accuracy being negatively correlated with cost (Macalino et al., 2015; Sliwoski et al., 2014; Leelananda and Lindert, 2016).

In recent years, both LB QSAR models and SB docking vHTS have come-of-age as powerful tools for small molecule hit discovery (Geanes et al., 2016; Butkiewicz et al., 2013; Stein et al., 2020). Advances in molecular mechanics methods such as free energy perturbation (FEP) and thermodynamic integration (TI) have led to unprecedented in silico rank-ordering of scaffold derivatives during hit-to-lead optimization (Wang et al., 2019a, 2015; Zou et al., 2019; Jorgensen and Thomas, 2008). Ongoing investigations in machine learning (ML) and quantum chemistry are poised to increase the predictive power of our CADD score functions (Lu et al., 2019; Brown et al., 2021; Kirkpatrick et al., 2021). Emerging strategies leverage principles from ML methods developed for LB CADD with physics-based methods developed for SB CADD (Gentile et al., 2022, 2020).

Traditionally, LB and SB CADD methods have been employed to perform vHTS. More recently, however, a number of algorithms have emerged that enable on-the-fly drug design. Some of these are tantamount

to vHTS and leverage the one-shot synthetic accessibility of made-on-demand libraries to propose efficient routes for molecular design (Bellmann et al., 2022; Schmidt et al., 2021; Sadybekov et al., 2022). Compared with other drug design approaches, the one-shot made-on-demand strategy greatly increases the synthetic throughput and synthesizability of candidate compounds. Other algorithms leverage ML, combinatorial chemistry, and/or multi-component reaction-based design to generate small molecule libraries with favorable predicted properties and activities (Popova et al.; Zhavoronkov et al., 2019; Brown et al., 2022). Combined with novel approaches aimed at improving the accuracy of predictions for physicochemical properties, such as solubility (Boobier et al., 2020), these methods have the potential to accelerate the drug discovery process.

All of these methods represent important advances; however, they exist largely in isolation as highly specialized protocols. CADD requires adaptability. The nature and scope of a CADD challenge is heavily influenced by factors such as the availability of training data, knowledge of the target chemical space, the presence (or absence) of experimental characterization of the drug target and putative binding pocket(s), the flexibility (dynamics) of the target, the size of the system under investigation, the expected accuracy of the score function in the given system, and more. While specialized tools can be highly valuable in certain circumstances, they may be of limited utility in others.

Indeed, there remains substantial attrition in the development of a compound from lead to FDA-approved therapy (Moreno and Pearson, 2013; Waring et al., 2015). The primary causes for these failures in clinical trials are lack of efficacy or safety (Harrison, 2016). In oncology specifically, kinases are a frequent drug target, and toxicity due to off-target effects is widely appreciated (Klaeger et al., 2017; Lin et al., 2019). Thus, in order to increase the success of candidate drugs in clinical trials, it is critical to develop new CADD technologies that directly address these limitations and are extensible to future challenges.

This dissertation is thematically separated into two major components. First, it describes novel mechanisms of oncogenic activation and therapeutic resistance in human epidermal growth factor receptors 1 (EGFR) and 2 (HER2), demonstrating in the process how mutation-induced changes in protein conformational free energy landscapes require innovative solutions in drug design. Second, it details the development of a new framework for small molecule drug design that integrates the BioChemical Library (BCL) cheminformatics toolkit with the Rosetta macromolecular modeling software suite. The new drug design framework is built specifically to address difficulties involved in designing small molecules to bind to dynamic proteins, such as EGFR kinase.

Chapter 2 describes the mechanism of action of the G724S resistance mutation in EGFR, which emerges in some non-small cell lung cancer (NSCLC) oncogenic variants as a response to first-line treatment with the third-generation tyrosine kinase inhibitor (TKI) osimertinib. Portions of this chapter are taken from Brown, B. P.*; Zhang, Y.-K.*; Westover, D.; Yan, Y.; Qiao, H.; Huang, V.; Du, Z.; Smith, J. A.; Ross, J. S.; Miller,

V. A.; Ali, S.; Bazhenova, L.; Schrock, A. B.; Meiler, J.; Lovly, C. M. On-Target Resistance to the Mutant-Selective EGFR Inhibitor Osimertinib Can Develop in an Allele-Specific Manner Dependent on the Original EGFR-Activating Mutation. Clin. Cancer. Res. 2019, 25 (11), 3341–335135.

Chapter 3 demonstrates that EGFR Ex19Del variants are a heterogeneous class of oncogenic mutants whose activation and TKI sensitivity are dictated by unique conformational preferences and catalytic activity. Portions of this chapter are in review for publication. This chapter is a collaborative work of Benjamin P. Brown*, Yun-Kai Zhang*, Soyeon Kim*, Yingjun Yan, Zhenfang Du, Jiyoon Kim, Abigail Leigh Hartzler, Michele L. LeNoue-Newton, Adam W. Smith, Jens Meiler, and Christine M. Lovly (*These authors contributed equally).

Chapter 4 describes the mechanistic basis of oncogenic activation for a new class of EGFR variants in NSCLC – kinase domain duplications (KDD). It also discusses the role of EGFR-KDD linker dynamics in promoting enhanced activity relative to wild-type. Portions of this chapter are taken from Du, Z.*; Brown, B. P.*; Kim, S.; Ferguson, D.; Pavlick, D. C.; Jayakumaran, G.; Benayed, R.; Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M.; Ali, S. M.; Schrock, A. B.; Zehir, A.; Ladanyi, M.; Smith, A. W.; Meiler, J.; Lovly, C. M. Structure–Function Analysis of Oncogenic EGFR Kinase Domain Duplication Reveals Insights into Activation and a Potential Approach for Therapeutic Targeting. Nature Communications 2021, 12 (1), 138236.

Chapter 5 explores therapeutic strategies for breast cancer involving various HER2 oncogenic mutations. It also provides a mechanistic explanation for the preferential co-mutation of HER3 E928G with specific HER2 kinase domain mutations. Portions of this chapter are taken from Hanker, A. B.*; Brown, B. P.*; Meiler, J.*; Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; Sheehan, J. H.; He, J.; Lalani, A. S.; Arteaga, C. L. Co-Occurring Gain-of-Function Mutations in HER2 and HER3 Modulate HER2/HER3 Activation, Oncogenesis, and HER2 Inhibitor Sensitivity. Cancer Cell 2021, 39 (8), 1099-1114.e837.

Collectively, Chapters 2 – 5 demonstrate that mutation-induced changes in conformational equilibrium can be responsible for profound alterations in protein-protein dimerization propensities, enzymatic activity, and sensitivity and resistance to TKIs. They identify areas for improvement in our current standard-of-care treatments for patients with NSCLC and breast cancer. Importantly, they highlight the need for software that is capable of simulating drug design while accounting for large conformational transitions, protein sequence changes, and other complex system-specific challenges.

Our approach for creating a modular, customizable drug design platform requires several algorithmic advancements. Chapter 6 introduces a new flexible, property-based small molecule flexible alignment algorithm in the BCL. The algorithm combines a customizable chemical property distance metric with efficient

alignment co-space sampling moves to identify alignments. Portions of this chapter are taken from Brown, B. P.; Mendenhall, J.; Meiler, J. BCL::MolAlign: Three-Dimensional Small Molecule Alignment for Pharmacophore Mapping. J. Chem. Inf. Model. 2019, 59 (2), 689–70138.

Chapter 7 describes a novel approach for rapid, interpretable SB scoring of protein-ligand interactions using deep neural networks (DNN). The score function is target agnostic and minimizes ligand bias by only utilizing protein-ligand atomic property correlations discretized into signed distance bins. Portions of this chapter are taken from Brown, B. P.; Mendenhall, J.; Geanes, A. R.; Meiler, J. General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps. J. Chem. Inf. Model. 2021, 61 (2), 603–62017.

Chapter 8 illustrates the new drug design framework, which in addition to the components in Chapters 6 and 7 also includes a series of chemical perturbations, a mutable atom selection module, and internal druglikeness filters. This chapter also discusses the integration of the BCL into Rosetta to enable protocol development that also makes use of Rosetta's extensive array of macromolecular modeling tools. Portions of this chapter are in review for publication. This chapter is a collaborative work of Benjamin P. Brown, Jeffrey Mendenhall, Rocco Moretti, Sergey Lyskov, Alexander R. Geanes, Darwin Fu, Sandeep Kothiwale, Edward W. Lowe, and Jens Meiler.

Chapter 9 summarizes the collective works of Chapters 2 – 8 and identifies ongoing and future directions.

CHAPTER 2

On-target resistance to the mutant-selective EGFR inhibitor osimertinib can develop in an allele specific manner dependent on the original EGFR activating mutation

This chapter is taken from Brown, B. P.*; Zhang, Y.-K.*; Westover, D.; Yan, Y.; Qiao, H.; Huang, V.; Du, Z.; Smith, J. A.; Ross, J. S.; Miller, V. A.; Ali, S.; Bazhenova, L.; Schrock, A. B.; Meiler, J.; Lovly, C. M. Clin. Cancer. Res. 2019, 25 (11), 3341–335135 (*These authors contributed equally).

2.1 Introduction

Oncogenic mutations in the EGFR tyrosine kinase domain are found in 15-30% of non-small cell lung carcinomas (NSCLC) (Lynch et al., 2004; Pao et al., 2004). Of these cases, approximately 90% can be attributed to in-frame deletions within exon 19 (Ex19Del) or missense mutations in exon 21 (L858R), which occur with approximately equal prevalence (Lynch et al., 2004; Pao et al., 2004). Multiple phase III clinical trials have shown that patients with EGFR-mutant tumors experience $>$70% radiographic response rates (RRs) and a statistically significant improvement in progression-free survival (PFS) when treated with first-generation (erlotinib, gefitinib) or second-generation (afatinib) EGFR tyrosine kinase inhibitors (TKIs) as compared with platinum based chemotherapy (Sequist et al., 2013; Rosell et al., 2012; Mitsudomi et al., 2010; Maemondo et al., 2010). However, response to these targeted agents is transient, and acquired therapeutic resistance typically develops within 8-10 months. In approximately 60% of cases, resistance is acquired through acquisition of a secondary EGFR mutation, EGFR T790M (Oxnard et al., 2018; Stewart et al., 2015; Yu et al., 2013). Osimertinib, a mutant-selective third-generation covalent inhibitor, was developed specifically to target T790M. For these reasons, the clinical standard of care for EGFR-mutant NSCLC has been treatment with first or second generation TKIs followed by treatment with osimertinib post-progression on first line therapy (Yang et al., 2017). Recently, osimertinib became approved as first-line therapy (Soria et al., 2018).

Unfortunately, resistance mutations may also emerge against osimertinib therapy (Papadimitrakopoulou VA, 2018; Ramalingam SS, 2018). The most well described to date is C797S, which is detected in approximately 10%-19% of patients with first-line and second-line osimertinib resistance (Piotrowska et al., 2018; Ramalingam et al., 2018). Mutation of C797 to serine prevents covalent adduct formation between osimertinib and the EGFR kinase domain (Thress et al., 2015; Yosaatmadja et al., 2015). We (Oztan et al., 2017) and others (Piotrowska et al., 2018; Peled et al., 2017; Fassunke et al., 2018) have also identified G724S as a mutation which is selected for in osimertinib resistant tumors. Unlike C797S, G724S was not predicted based on in vitro studies (Yu et al., 2007; Ercan et al., 2015), and the precise mechanism whereby G724S mutation

confers osimertinib resistance is unknown.

The most fundamental principle of structural biology is that sequence determines structure and structure determines function. To determine the relationship between classical EGFR kinase activating mutations (Ex19Del and L858R), acquired G724S mutation, and osimertinib resistance, we employed an integrated computational / experimental approach. Our results suggest that G724S is a resistance mutation that develops with Ex19Del but not L858R and provide mechanistic insight into this process at the structural level.

2.2 Results

2.2.1 A G724S-mediated conformational change in the glycine-rich P-loop reduces binding affinity of osimertinib to Ex19Del/G724S but not to L858R/G724S

To determine the structural effects of G724S mutation on osimertinib binding, we performed a series of Gaussian accelerated molecular dynamics (GaMD) simulations (Miao and McCammon, 2017; Miao et al., 2015) of wild-type EGFR (WT), Ex19Del (unless otherwise stated, the canonical variant E746_A750del), Ex19Del/G724S, L858R, and L858R/G724S in the drug-unbound (apo) state. Analysis of our initial simulations suggests G724S may increase P-loop backbone conformation fluctuations (Figure 2.1). These data are intriguing because EGFR has previously been shown to bind osimertinib with a characteristic "bent" P-loop conformation (Yosaatmadja et al., 2015), and we hypothesized that G724S could reduce osimertinib binding through disruption of the bent P-loop conformation. Previous literature on protein conformational dynamics has cautioned against inferring functional mechanisms from RMSF statistics alone (Farmer et al., 2017). Therefore, to test our hypothesis, we performed GaMD simulations of Ex19Del, Ex19Del/G724S, L858R, and L858R/G724S reversibly bound with osimertinib. We similarly examined these four mutants with the second-generation, wild-type selective EGFR TKI, afatinib, as a control. Afatinib was selected as a control in our study for multiple reasons. Afatinib has previously been reported to be a potential therapeutic agent in the setting of Ex19Del/G724S-mediated NSCLC based on a patient case report (Oztan et al., 2017). Similar to osimertinib, afatinib is an irreversible EGFR inhibitor that has received regulatory approval for treatment of EGFR-mutant lung cancer.

Osimertinib and afatinib both irreversibly bind EGFR through covalent adduct formation. In order to form an irreversible complex, they must first form a reversible, non-covalent complex (Figure 2.2A). Disruption of the reversible complex formation is expected to reduce formation of adduct. A previously determined crystallographic structure of EGFR kinase reversibly bound to osimertinib demonstrates that osimertinib binding is accommodated through a well-defined "bent" P-loop conformation (Figure 2.2B) (Yosaatmadja et al., 2015). This bent P-loop conformation allows the F723 phenyl ring to make an energetically favorable contact with the indole ring of osimertinib, contributing to its affinity (Yosaatmadja et al., 2015).

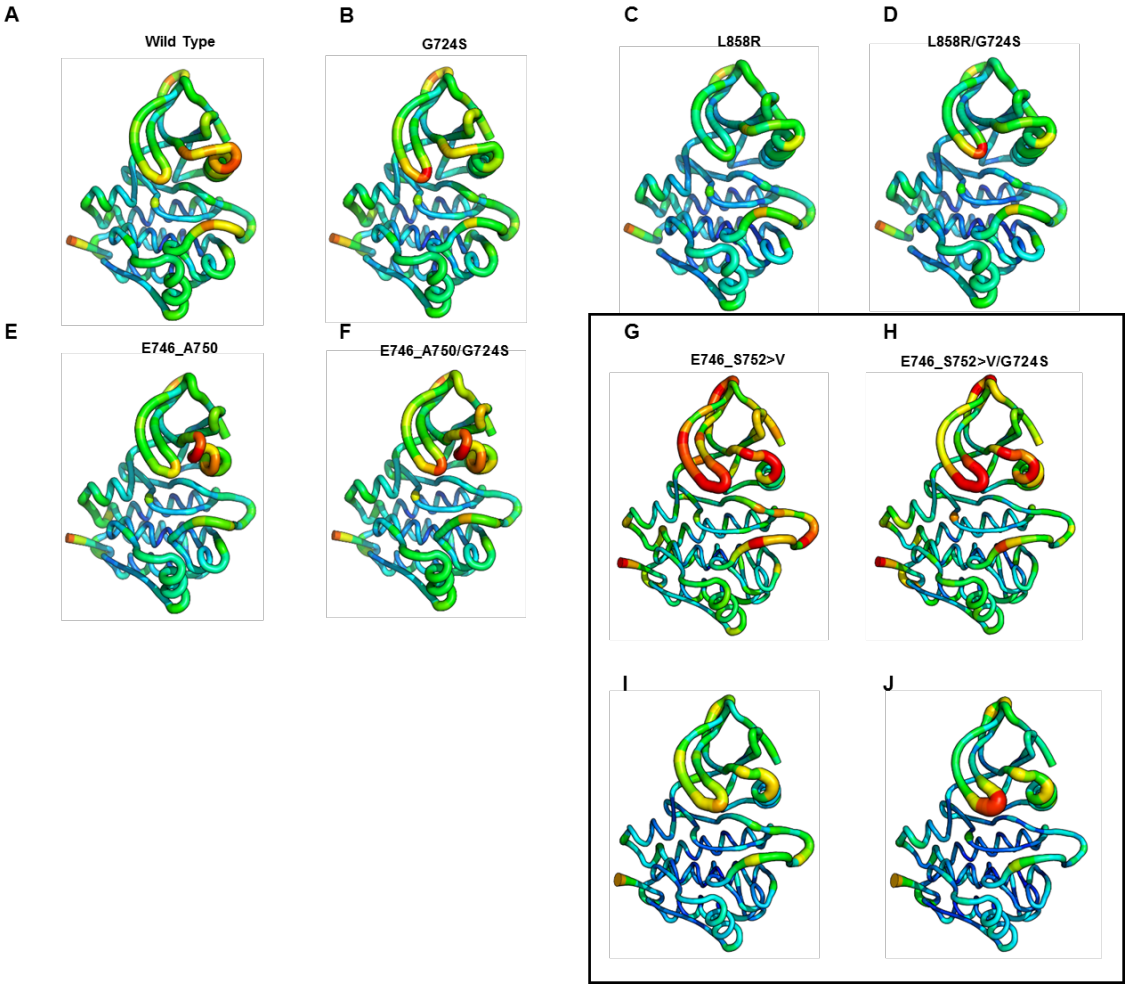


Figure 2.1: G724S increases P-loop backbone fluctuations. We performed 500 ns GaMD simulations of EGFR (A) WT, (B) G724S, (C) L858R, (D) L858R/G724S, (E) E746_A750, (F) E746_A750/G724S, (G and I) E746_S752>V, and (H and J) E746_S752>V/G724S. Per-residue RMSF is scaled between 0 (green) and 3 (red) $\mathring{A}$ (A-H) or 0 and 5 $\mathring{A}$ (I-J).

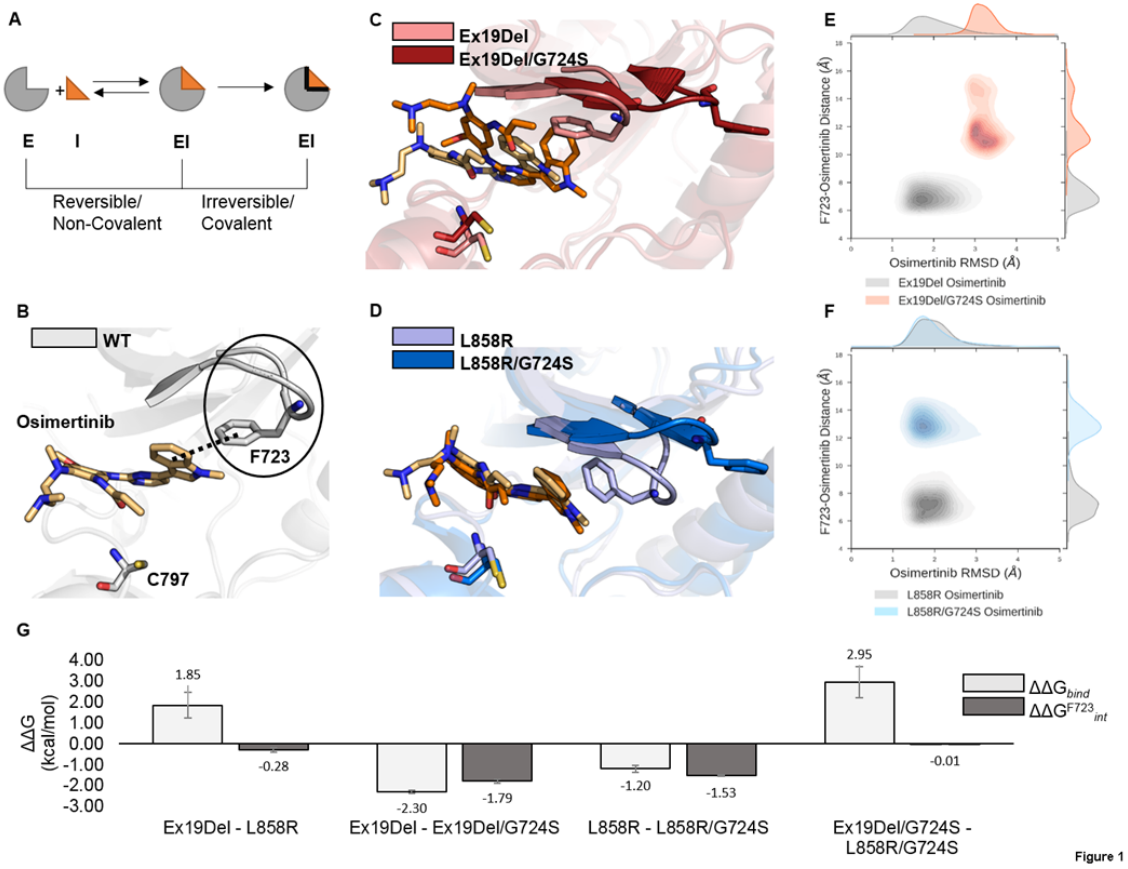


Figure 2.2: Stability of osimertinib in reversible complexes with EGFR mutants. EGFR mutants reversibly bound to osimertinib were simulated with GaMD. A schematic representation of a simplified binding equilibrium for a covalently-binding inhibitor is depicted such that E = Enzyme target, I = Inhibitor, and EI = Enzyme-Inhibitor complex (A). Each simulation was performed in triplicate for a total of 12 independent 250 ns GaMD simulations. Representative images of osimertinib reversibly bound to WT (PDB ID 4ZAU; the solid black line indicates the bent P-loop; the dashed black line indicates the contact between the F723 phenyl and osimertinib indole ring; (B), Ex19Del and Ex19Del/G724S (C), and L858R and L858R/G724S (D) are displayed. Trajectory frames were extracted every 10 ps and plotted as osimertinib RMSD from the equilibrated start structure (x-axis) and distance between the phenyl ring of F723 and the indole ring of osimertinib (y-axis; E – F). RMSD vs. distance plots include data from 3 independent trajectories for each mutant – inhibitor pair (E – F). Select relative osimertinib binding free energies are plotted as averages across 3 independent trajectories; error bars indicate standard error of the mean (G).

$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{solv} - T\Delta S$

$\Delta G_{F723int} = \Delta E_{MM} + \Delta G_{solv}$

$\Delta\Delta G = \Delta G1 - \Delta G2$

Our GaMD simulations illustrate that G724S rigidifies the tip of the P-loop by stabilizing a β-bend conformation (Figure 2.2C, D; Figure 2.3).

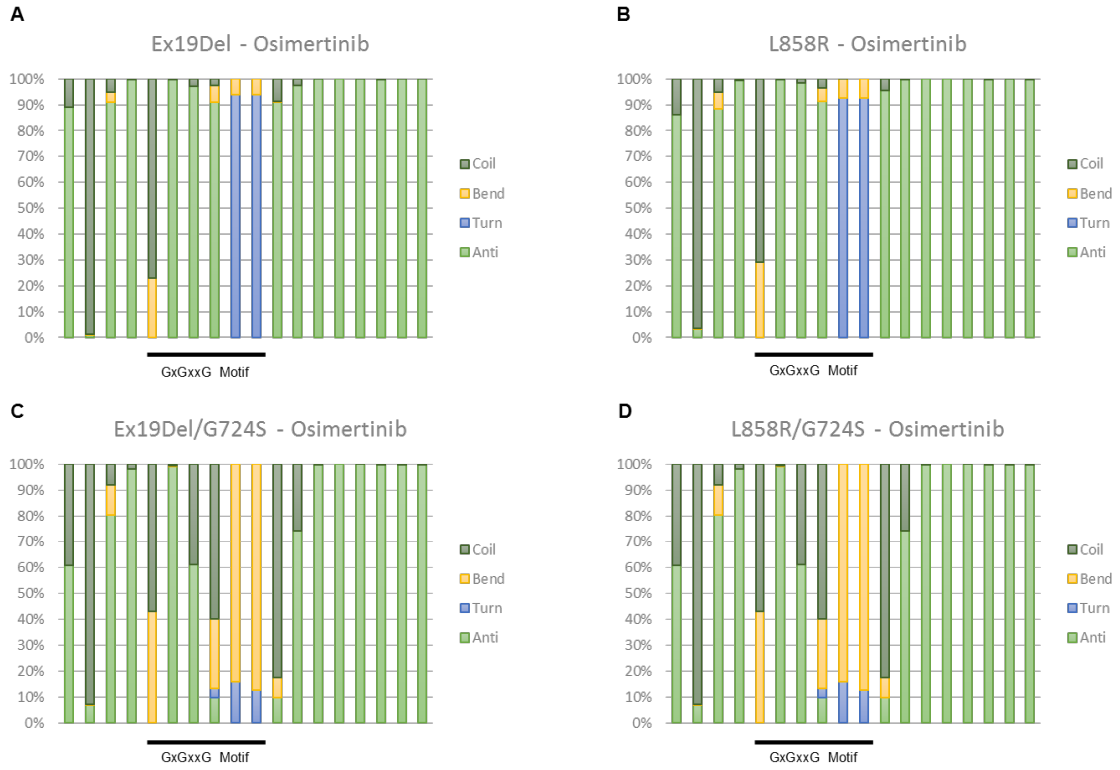


Figure 2.3: G724S induces an α-turn to β-bend conformational shift in the P-loop.

As a result, Ex19Del/G724S and L858R/G724S cannot form a stable bent P-loop conformation when bound to osimertinib. The rigidified P-loop displaces F723 from contact with osimertinib (Figure 2.2C – F). Interestingly, however, we found evidence of reduced stability of the osimertinib-bound Ex19Del/G724S complex but not the osimertinib-bound L858R/G724S complex. In our simulations, osimertinib maintains an RMSD of 1 – 2 Å from its native binding pose in Ex19Del and L858R. Displacement of F723 from contact with osimertinib is associated with an increase in osimertinib RMSD to 3 – 4 Å in Ex19Del/G724S but not in L858R/G724S (Figure 2.2E, F). In contrast, afatinib forms a stable reversible complex in all four cases (Ex19del, Ex19del/G724S, L858R, and L858R/G724S) (Figure 2.4). These models suggest that structural perturbations from G724S, which disrupt binding of osimertinib, fail to notably effect binding of afatinib. These data support a potential role for afatinib in treating patients with G724S.

To further investigate these differences, we applied the molecular mechanics-generalized Born surface area method (MM/GBSA) to compute the binding free energies of osimertinib with Ex19Del, Ex19Del/G724S, L858R, and L858R/G724S. Our calculations predict a 2.3 kcal/mol reduction in osimertinib binding free energy ($\Delta\Delta G_{bind}$) with Ex19Del/G724S (Figure 2.2G). Our binding free energy calculations also suggest that

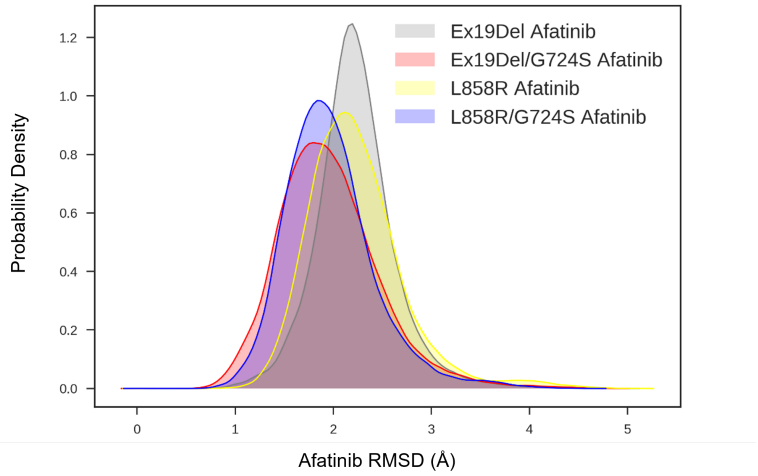


Figure 2.4: Afatinib forms a stable reversible complex with EGFR independent of G724S status.

osimertinib reversibly binds L858R more tightly than Ex19Del by 1.9 kcal/mol. (Figure 2.2G). The osimertinib binding free energies computed for Ex19Del and L858R/G724S are indistinguishable, within error, suggesting that the reduction in binding affinity accompanying the addition of the G724S mutation in L858R should not confer osimertinib resistance.

In addition, energy decomposition analysis supports our qualitative observation that F723 contributes favorably to osimertinib binding in both Ex19Del and L858R (the interaction energy of F723, $\Delta GF723_{int}$, defined $\Delta GF723_{int} = \Delta E_{MM} + \Delta G_{solv}$, is approximately -1.8 and -1.5 kcal/mol, respectively), and that addition of G724S prevents this interaction (Figure 2.2G). As expected based on crystallographic evidence, our simulations show that F723 contributes considerably less to the interaction of EGFR with afatinib (Solca et al., 2012). Consistent with Fassunke et al. (Fassunke et al., 2018), our afatinib relative binding free energies are less affected by G724S versus osimertinib. Altogether, these data suggest G724S may function as a resistance mutation to osimertinib in Ex19Del/G724S, but not in L858R/G724S.

2.2.2 In vitro expression of Ex19Del/G724S, but not L858R/G724S, is associated with osimertinib resistance

To test our simulation predictions, we first examined the ability of osimertinib to inhibit EGFR autophosphorylation of various EGFR single, double, and triple mutants. Of note, to date, G724S has been detected in both the absence and presence of T790M (Oztan et al., 2017; Peled et al., 2017). Therefore, we modeled all possibilities in our experimental studies. Osimertinib was effective at inhibiting EGFR autophosphorylation in 293FT cells expressing Ex19Del and Ex19Del/T790M, but not in 293FT cells expressing

Ex19Del/C797S or Ex19Del/T790M/C797S, as C797S mutation has previously been associated with osimertinib resistance(16,27) (Figure 2A). Likewise, osimertinib was ineffective at blocking autophosphorylation of EGFR Ex19Del/G724S and Ex19Del/T790M/G724S mutants.

We also tested the efficacy of osimertinib against L858R variant combinations. Analogous to the Ex19Del data above, phosphorylation of L858R and L858R/T790M were inhibited by osimertinib while C797S containing variants (L858R/C797S and L858R/T790M/C797S) were insensitive to this agent (Figure 2B). In contrast to the Ex19Del variant data, phosphorylation of L858R/G724S and L858R/T790M/G724S were potently inhibited by osimertinib (Figure 2B). These data are consistent with our simulations, which suggested a difference in the drug binding properties between Ex19Del and L858R when combined with G724S mutation. Altogether, these data suggest that G724S functions as a resistance mutation in the context of Ex19Del but not L858R.

Next, we attempted to define strategies to overcome osimertinib resistance mediated by G724S mutation. In particular, we focused on the efficacy of earlier generations of wild-type selective EGFR TKIs. Previous studies have demonstrated that C797S-containing EGFR variants, which are resistant to osimertinib, retain sensitivity to the first generation EGFR TKIs (erlotinib, gefitinib) (16). We sought similar strategies for G724S-containing EGFR variants. We quantitatively evaluated several TKIs on Ex19Del-series mutants by stably transducing Ex19Del EGFR variants into Ba/F3 cells and measuring IL-3-independent growth at multiple inhibitor concentrations (Figure 2.5 C-E). As expected, growth of cells expressing EGFR Ex19del/C797S and EGFR Ex19del/G724S was insensitive to osimertinib. Cell lines expressing Ex19del/C797S and Ex19del/G724S were also cross-resistant to another mutant-selective EGFR-TKI, rociletinib (Figure 2.6).

In accord with previous data (Thress et al., 2015), cells expressing Ex19Del/C797S were sensitive to the effects of the first generation EGFR TKI, erlotinib, with an EC50 paralleling that of the original Ex19Del single mutant (16.12 nM vs. 13.71 nM, respectively, Figure 2D). However, the Ex19Del/G724S mutant was insensitive to the effects of erlotinib (EC50 $> 1\ \mu$M). Our structural data suggested that afatinib may retain efficacy against the Ex19Del/G724S double mutant (Figure 2.4). In accord with these data, the growth of cells expressing this double mutant was inhibited with an EC50 of 29.63 nM afatinib (Figure 2E). Likewise, autophosphorylation of the Ex19Del/G724S in stably transduced NR6 cells was potently inhibited by afatinib, but not erlotinib or osimertinib (Figure 2F, Figure 2.7), while the autophosphorylation of the L858R/G724S was potently inhibited by both afatinib and osimertinib (Figure 2G, Figure 2.7).

Importantly, previous in vitro screens failed to identify G724S as a resistance mutation (Yu et al., 2007; Ercan et al., 2015). Our data suggest that this may be because these screens generated missense mutants beginning with WT, L858R, or L858R/T790M. Our data suggest that G724S functions as a resistance mutation in the context of Ex19Del but not L858R. Moreover, our results provide additional evidence that afatinib, but

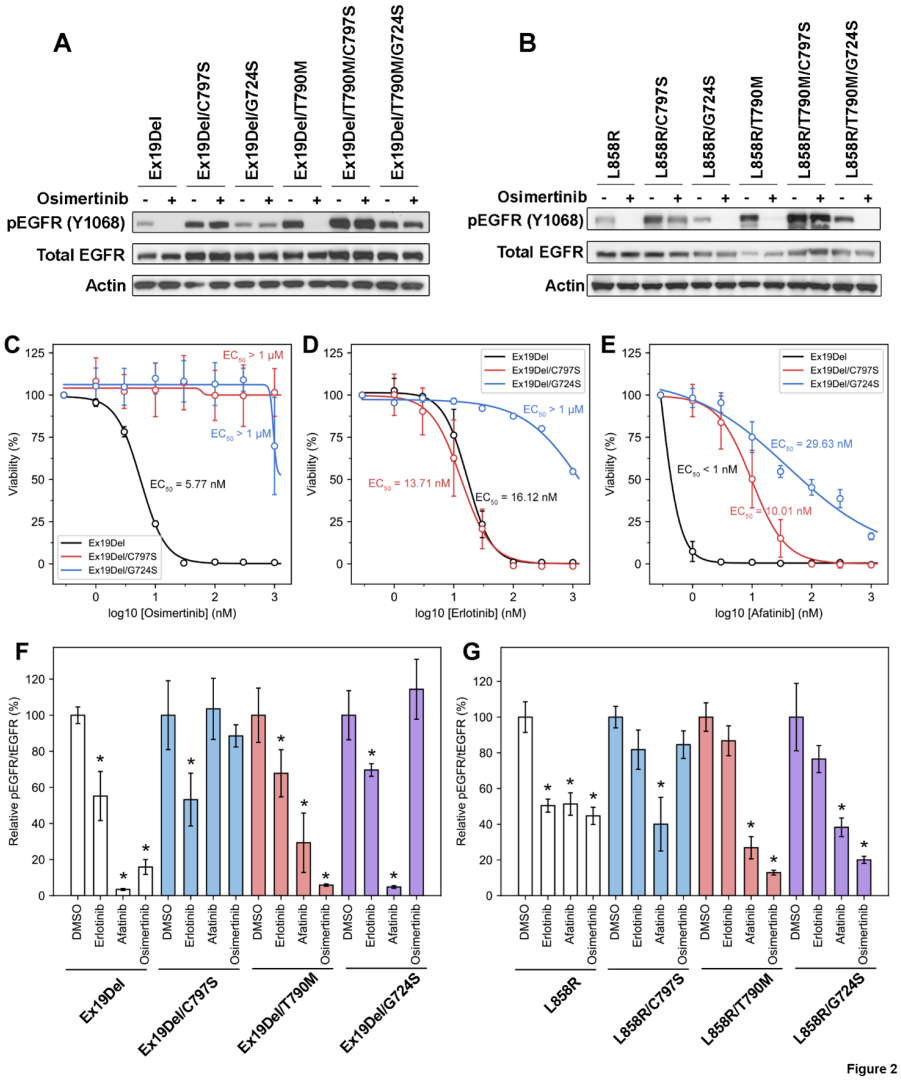


Figure 2.5: EGFR G724S mediates osimertinib resistance in EGFR Ex19Del but not EGFR L858R mutants. (A) 293FT cell transduced with different EGFR del19 variants were treated with 100 nM osimertinib for 4 hours. Cellular lysates were probed with the indicated antibodies. (B) 293FT cell transduced with different EGFR L858R variants were treated with 100 nM osimertinib for 4 hours. Cellular lysates were probed with the indicated antibodies. Ba/F3 EGFR Ex19Del, Ex19Del19/C979S, Ex19Del/G724S were treated with increasing amount of (C) osimertinib, (D) erlotinib or (E) afatinib for 72 hours. CellTiter Blue assays were performed to assess cell viability. Each point represents three replicates. Data are presented as the mean percentage of viable cells compared to control ± SD. NR6 cells transduced with (F) different EGFR del19 variants or (G) different EGFR L858R variants were treated with either DMSO, 100 nM erlotinib, 100 nM afatinib, or 100 nM osimertinib for 4 hours. Relative pEGFR/tEGFR values are calculated by the density of pEGFR signal divided by the density of tEGFR signal, then normalized by the DMSO-treated group in each cell line. Density of western blots was analyzed by ImageJ. *: $p < 0.05$ as compared to DMSO-treated group in each cell line. Data and illustrations for this figure produced by Zhang, Y.-K., Westover, D.; Yan, Y.; Qiao, H.; Huang, V.; Du, Z., and Lovly, C.M.

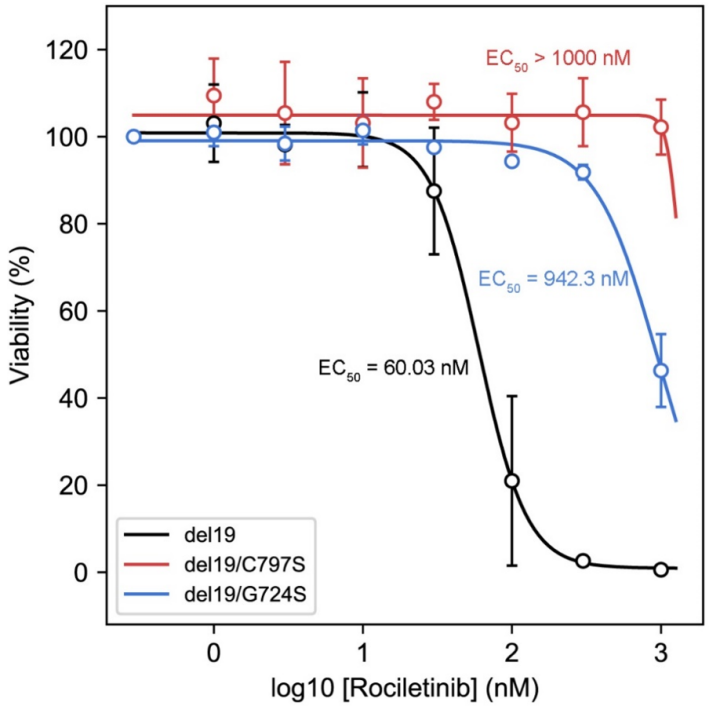


Figure 2.6: Efficacy of Rociletinib against EGFR Ex19Del containing variants. Data and illustrations for this figure produced by Zhang, Y.-K., Westover, D.; Yan, Y.; Qiao, H.; Huang, V.; Du, Z., and Lovly, C.M.

not osimertinib or erlotinib, can function effectively as an inhibitor of Ex19Del/G724S.

2.2.3 G724S emerges as a resistance mutation in Ex19Del but not L858R-mediated NSCLC

To date, four independent reports (Piotrowska et al., 2018; Oztan et al., 2017; Peled et al., 2017; Fassunke et al., 2018) have identified G724S as an emergent mutation in patients who have developed acquired resistance to osimertinib, with the frequency of G724S being 13% (higher than the frequency of C797S) in a recent paper by Fassunke and colleagues (Fassunke et al., 2018). Interestingly, all of these patients harbored Ex19Del as the original activating mutation (Piotrowska et al., 2018; Oztan et al., 2017; Peled et al., 2017; Fassunke et al., 2018). Our computational and experimental data suggest that G724S confers resistance to osimertinib in Ex19Del but not L858R; nevertheless, it is possible that L858R/G724S exists in a subset of EGFR-mutant NSCLC patients. To investigate the prevalence of EGFR G724S mutation, we analyzed data from tissue and plasma DNA samples within the Foundation Medicine database. Consistent with our computational and experimental evidence, G724S co-occurred with an Ex19Del variant in 15/19 cases, and L858R/G724S was not identified (Figure 2.8A). Given that the likelihood of observing Ex19Del versus L858R in EGFR-mutant NSCLC is approximately equal (Zhang et al., 2016), it is exceedingly unlikely that

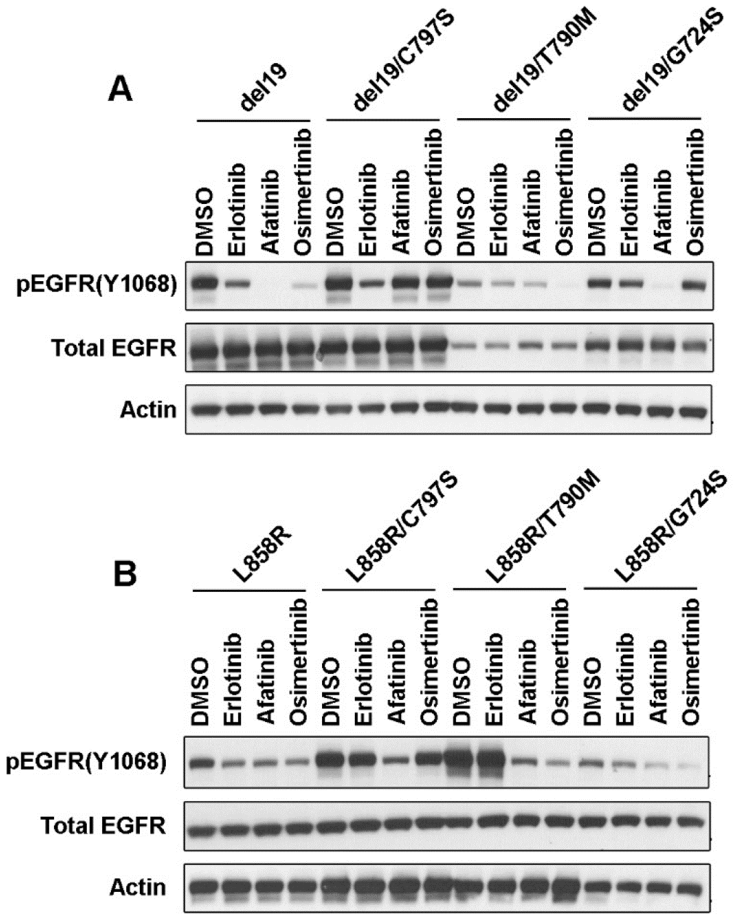


Figure 2.7: TKI inhibition profile of autophosphorylation against EGFR Ex19Del and L858R containing variants. Data and illustrations for this figure produced by Zhang, Y.-K., Westover, D.; Yan, Y.; Qiao, H.; Huang, V.; Du, Z., and Lovly, C.M.

L858R activating mutation would not be found in any of our patient samples without an additional bias.

In four cases (all Ex19Del variants), we were able to obtain tissue genomic profiling data at two unique time points. In three of these cases (Figure 2.8C-E), G724S allelic frequency is positively correlated with Ex19Del allelic frequency over time and decline of the T790M allele. Moreover, G724S is not present in the tumor biopsy from any of these four patients prior to Ex19Del; that is, the mutant allele frequency (MAF) of G724S starts at zero in all of these matched cases (Figure 2.8B – E). These data suggest that G724S emerges in a fraction of Ex19Del patients to promote disease progression.

To highlight one particular case (patient #15, Figure 2.8E), a 54 year old Caucasian gentleman never smoker was diagnosed with stage IV lung adenocarcinoma after presenting with abdominal pain. Tumor mutational testing was positive for an EGFR Ex19Del mutation. He was treated with first line erlotinib plus bevacizumab with partial response. Fifteen months after starting this combination therapy, he experienced progression of disease with enlargement of bilateral pulmonary nodules and a ground glass opacity in the left upper lobe. Repeat biopsy confirmed metastatic lung adenocarcinoma and tumor genetic testing at that time revealed the presence of EGFR Ex19Del and T790M mutations. He was thereafter treated with osimertinib and had a partial response lasting thirty months (Figure 2.8F). He experienced progression of disease with new metastases to the skull, liver, and bone. Tumor genetic testing of a repeat biopsy revealed the presence of EGFR Ex19Del, loss of T790M mutation, and gain of EGFR G724S mutation. He was treated with radiation therapy to the skull followed by systemic therapy with carboplatin and pemetrexed. Approximately four months after starting cytotoxic chemotherapy, he developed symptomatic pleural and pericardial effusions, which ultimately resulted in his demise.

Of note, G724S was also detected with the oncogenic missense mutant S768I in 2/19 cases, Shan and colleagues previously demonstrated that S768I stabilizes the active conformation by improving hydrophobic packing between the αC-helix and the β9-strand. G724S also occurred as an individual missense mutation in 2/19 cases (Figure 2.8A). The latter suggests that G724S could potentially be independently oncogenic. Indeed, G724S could support oncogenic growth of Ba/F3 cells (Figure 2.9A).

Of note, the G724S single mutant exhibits a TKI sensitivity profile very similar to Ex19Del in that this mutant can be effectively inhibited by erlotinib, afatinib, and osimertinib (Figure 2.10, Figure 2.9B). In addition, we identified nine cases of EGFR G724S as an isolated mutation in patients with small-cell lung carcinoma, bladder urothelial carcinoma, glioblastoma, breast cancer, and colorectal cancer. These data are consistent with recent evidence implicating G724S as an oncogenic driver in colorectal cancer (Cho et al., 2014) and suggest that patients with tumors harboring an isolated G724S mutation could be treated with FDA-approved EGFR TKIs, such as afatinib.

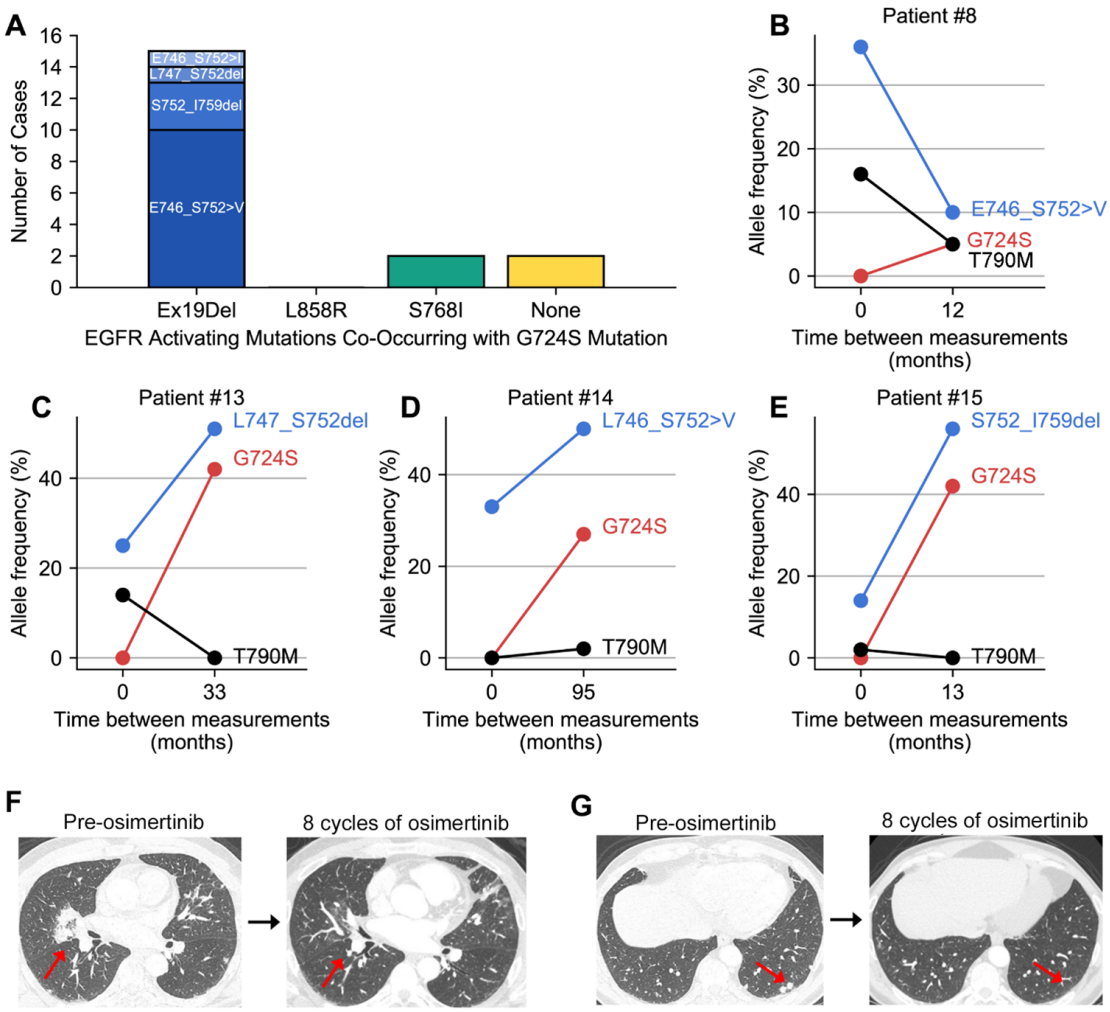


Figure 3

Figure 2.8: Prevalence of oncogenic EGFR mutations in NSCLC patient samples with G724S. (A) Bar chart depicting the number of cases of each oncogenic EGFR mutation associated with G724S in NSCLC patient samples with genomic profiling obtained through Foundation Medicine (total n=19). (B-E) Allelic frequencies for the specific Ex19Del variant, T790M, and G724S are plotted versus time between measurements for four cases for which tissue genomic profiling results were available at two independent time points. (F-G) Radiographic images for Patient 15 taken prior to osimertinib therapy (left) and after 8 cycles of osimertinib (right). The red arrows in the CT scan images show sites of disease that responded to osimertinib. Data and illustrations for this figure produced by Ross, J. S.; Miller, V. A.; Ali, S.; Bazhenova, L.; and Schrock, A. B.

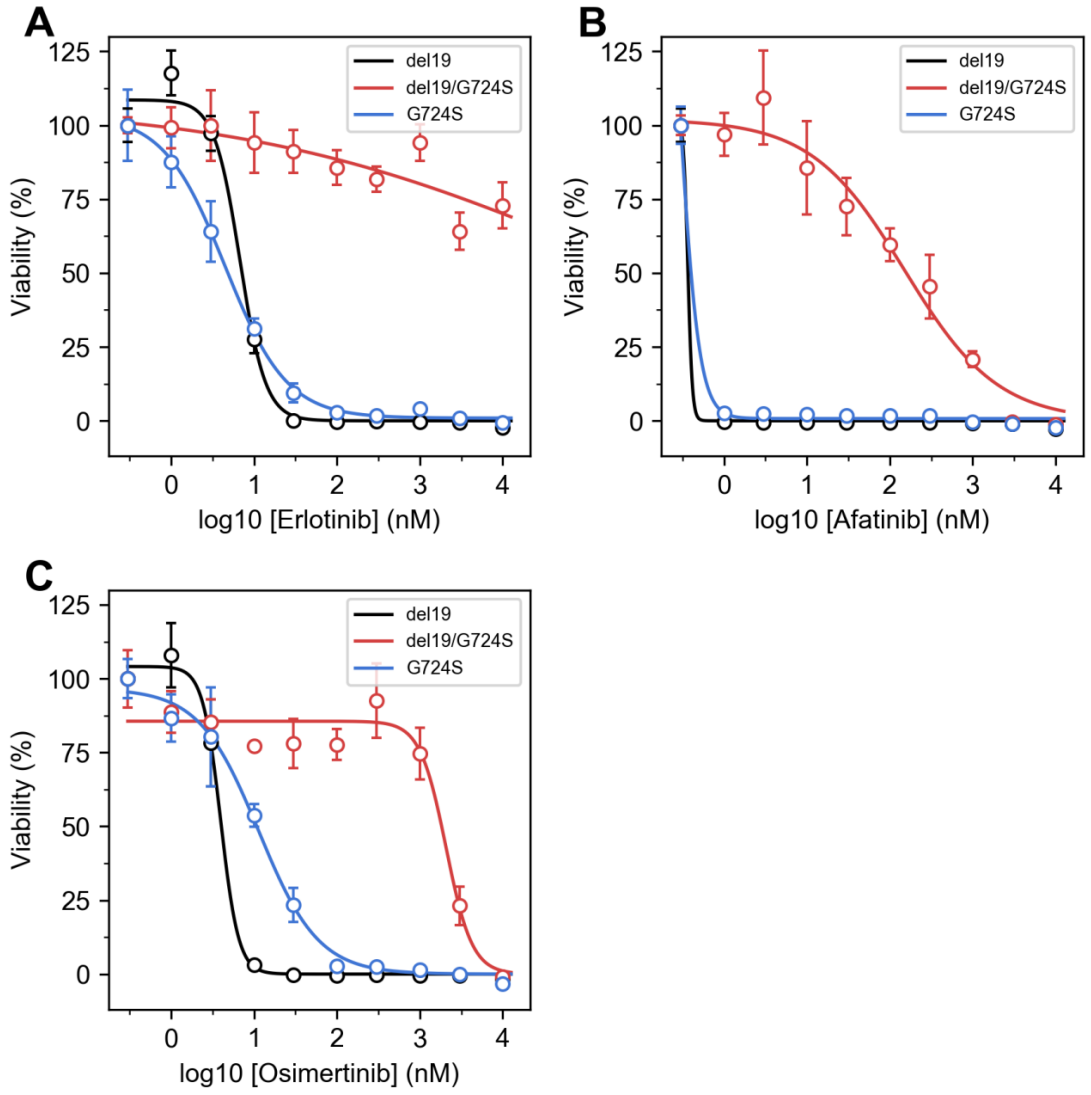


Figure 2.9: TKI inhibition profile of G724S, Ex19Del and Ex19Del/G724S. Data and illustrations for this figure produced by Zhang, Y.-K., Westover, D.; Yan, Y.; Qiao, H.; Huang, V.; Du, Z., and Lovly, C.M.

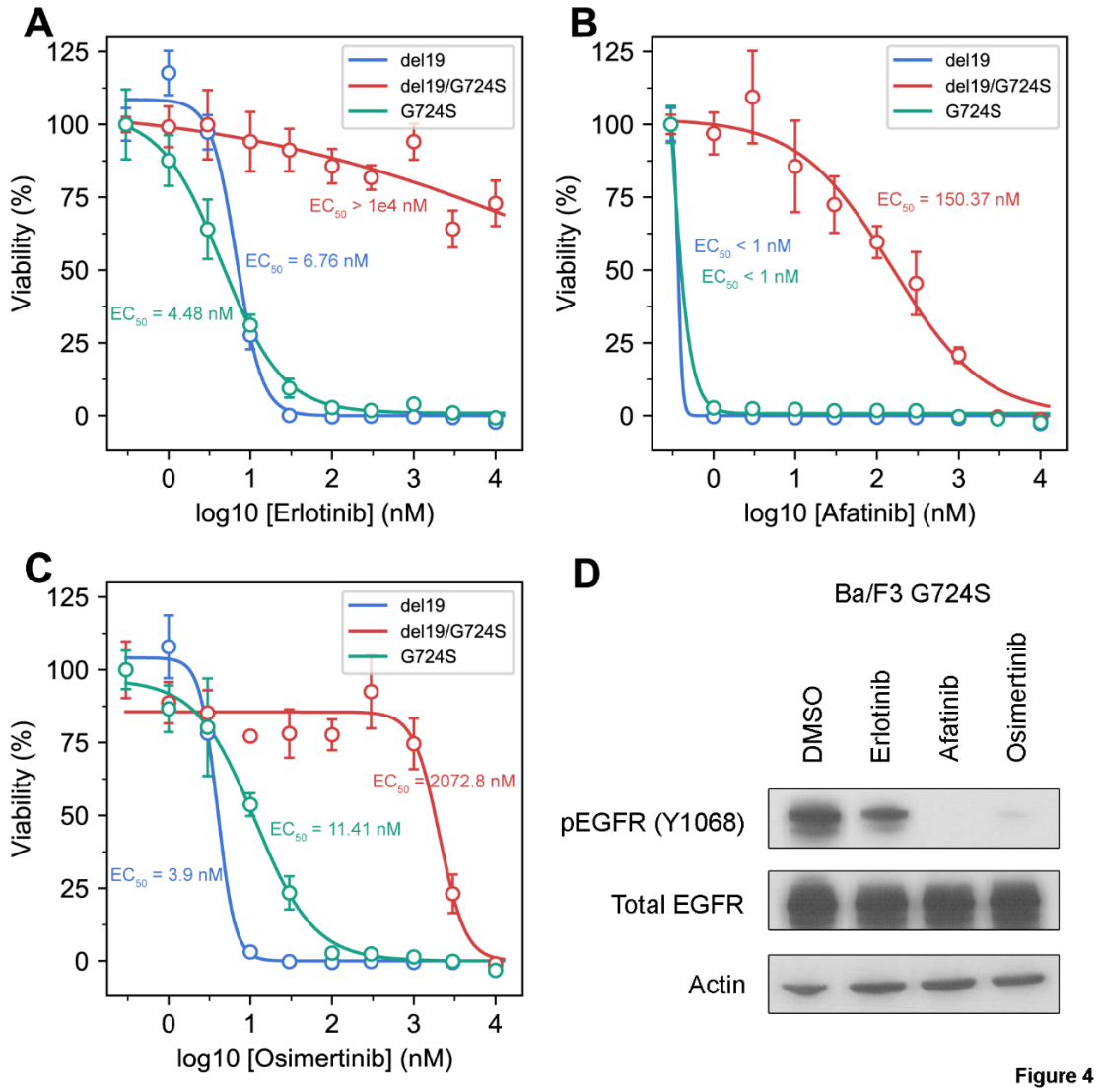


Figure 2.10: The EGFR G724S single mutant can be effectively inhibited by EGFR TKIs. Ba/F3 cells stably expressing EGFR Ex19Del, G724S, and Ex19Del/G724S were treated with increasing amounts of (A) erlotinib, (B) afatinib or (C) osimertinib for 72 hours. CellTiter Blue assays were performed to assess cell viability. Each point represents four replicates. Data are presented as the mean percentage of viable cells compared to control +/- SD. (D) Ba/F3 cells transduced with EGFR G724S were treated with either DMSO, 100 nM erlotinib, 100 nM afatinib, or 100 nM osimertinib for 4 hours. Cellular lysates were probed with the indicated antibodies. Data and illustrations for this figure produced by Zhang, Y.-K., Westover, D.; Yan, Y.; Qiao, H.; Huang, V.; Du, Z., and Lovly, C.M.

2.2.4 The catalytically active conformation of EGFR is better stabilized by E746_S752>VG724S than by E746_A750delG724S

Unexpectedly, all of the Ex19Del alterations co-occurring with G724S in patient tumor samples were rare variants. The Ex19Del variant occurring most frequently with G724S in this cohort was E746_S752>V (10/19), followed by S752_I759del (3/19), E746_S752>I (1/19), and L747_S752del (1/19). For context, approximately 67% of Ex19Del cases are attributed to the canonical variant, E746_A750del, while less than 2% are attributed E746_S752>V (Kobayashi and Mitsudomi, 2016). To better understand this enrichment in Ex19Del rare variants, we performed GaMD simulations for E746_S752>V and E746_S752>V/G724S in the apo-state and in reversible complex with osimertinib.

We utilized MM/GBSA to compute the relative binding free energies between the two sets of Ex19Del variants. The results displayed large statistical uncertainty in the calculation of the binding free energies, that we attribute to increased P-loop fluctuations in E746_S752>V and E746_S752>V/G724S relative to WT and the other variants (Figure 2.1). The majority of this difference is attributable to increased fluctuations in E746_S752>V, and just as in the cases of WT and E746_A750del the additional fluctuations associated with G724S in E746_S752>V occur primarily at the tip of the P-loop (Figure 2.1). Nevertheless, E746_S752>V, but not E746_S752>V/G724S, is able to stabilize a favorable contact between F723 and the indole ring of osimertinib, consistent with results obtained in the previous E746_A750del and E746_A750del/G724S osimertinib-binding simulations.

EGFR kinase activation is achieved through asymmetric dimerization of an acceptor EGFR kinase αC-helix with a donor kinase αH-helix. The acceptor kinase is the catalytically active dimer subunit (Zhang et al., 2006). In a seminal paper on EGFR dynamics, Shan and colleagues demonstrated that common oncogenic mutations increase activity by stabilizing the αC-helix inward conformation to promote asymmetric dimerization (Shan et al., 2012). We hypothesized that the unexpected enrichment of the E746_S752>V/G724S double mutant in clinical samples may result from increased stabilization of the αC-helix inward conformation in E746_S752>V/G724S relative to E746_S752>V. To test this hypothesis, we performed a detailed analysis of the conformational free energy landscape profiles of each EGFR variant in the apo-state.

Consistent with Shan and colleagues, results from our GaMD simulations of WT, E746_A750del, and L858R demonstrate increased stabilization of the αC-helix inward conformation compared to WT (Shan et al., 2012). Additionally, our simulations show that E746_S752>V stabilizes the αC-helix inward conformation relative to WT. Critically, our computational analyses suggest that E746_S752>V/G724S stabilizes the active αC-helix inward conformation even more than E746_S752>V (Figure 2.11E). In contrast, E746_A750del/G724S visits αC-in conformations less frequently than E746_A750del (Figure 2.11D). These

results suggest that E746_S752>V/G724S could lead to enhanced dimerization-dependent activation compared to E746_S752>V, while E746_A750del/G724S could lead to reduced dimerization-dependent activation compared to E746_A750del.

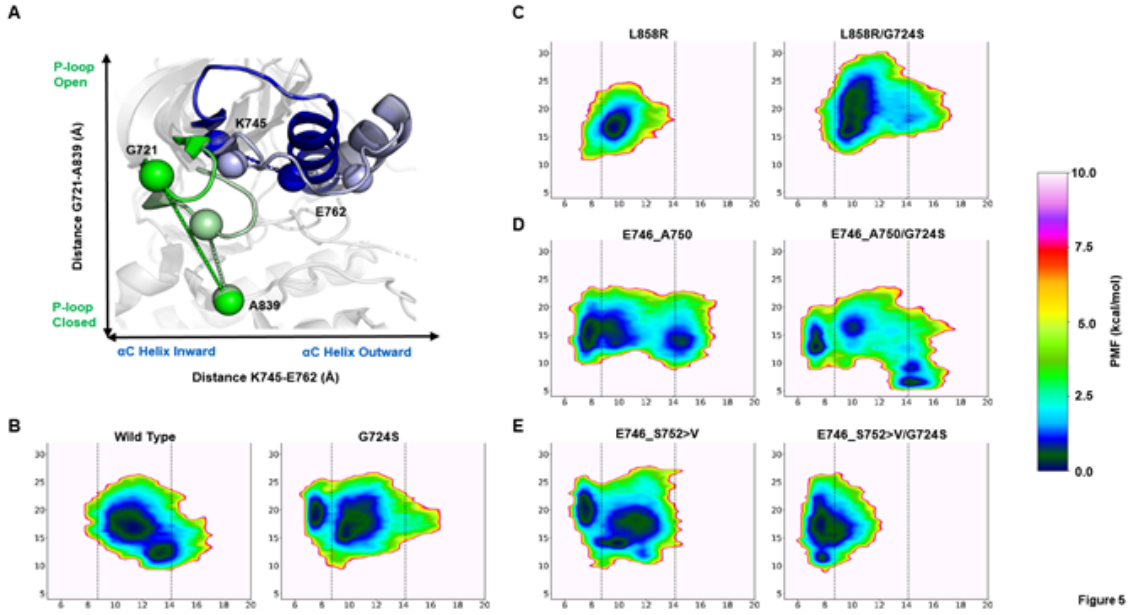


Figure 2.11: Conformational free energy landscape of EGFR kinase domain mutants. The reaction coordinate reference for the conformational free energy landscape of EGFR kinase mutants is indicated on a model of WT in the active (PDB ID 2ITX; bold colors) and inactive (PDB ID 3GT8; faded colors) conformations (A). Green spheres represent the distance (Å) between Hα1 of G721 and Cβ of A839. Blue spheres represent the distance between Cβ of K745 and Cβ of E762. The potential of mean force (PMF) with respect to the positions of the αC helix (x-axis) and P-loop (y-axis) are plotted for WT and G724S, L858R and L858R/G724S, E746_A750 and E746_A750/G724S, and E746_S752>V and E746_S752>V/G724S (B). The left and right vertical dashed lines on the free energy plots (C-E) indicate center-of-mass distances between K745 and E762 in active (PDB ID 2GS6) and inactive (PDB ID 2GS7) EGFR kinase, respectively. The left vertical dashed lined therefore represents the canonical EGFR kinase αC-helix inward conformation, while the right vertical dashed line represents the canonical EGFR kinase αC-helix outward conformation. All depicted simulations start from the active (αC-helix inward, activation loop outward) conformation. The energetic reweighting factor was approximated with cumulant expansion to the 2nd order. Free energy landscapes from the 500 ns GaMD simulations are depicted here.

Collectively, these data support G724S as a resistance mutation in Ex19Del over L858R, and that specific Ex19Del mutants may preferentially co-occur with G724S, potentially driven by differences in active conformation stability in the presence of G724S. In addition, our results suggest that G724S (as a single point mutation) also stabilizes the αC-helix inward conformation, consistent with reports that G724S may function as an oncogenic variant in colorectal cancer (Cho et al., 2014) (Figure 2.11A – D). Our data more broadly suggest that the underlying activating mutation profile of EGFR influences the development of drug resistance mutations. This has important implications for clinical management of patients with EGFR-mutant NSCLC.

2.3 Discussion

Notable advancements have been observed through the development of increasingly selective inhibitors of mutant oncoproteins (11). The design and clinical implementation of mutant-selective third generation EGFR TKIs, such as osimertinib, are an excellent example. Unfortunately, despite these advances, the development of resistance mutations to TKI therapy remains a significant barrier in attaining the best outcomes for patients with EGFR-mutant NSCLC. In addition to the previously identified C797S resistance mutation, our results demonstrate osimertinib resistance may emerge in the form of G724S mutations within the P-loop of the EGFR kinase domain. However, unlike C797S, our results also suggest that G724S-mediated resistance preferentially occurs in Ex19Del but not L858R. Indeed, extensive atomic-detail simulations at the structural level, multiple independent in vitro models, and patient genomic profiling all demonstrate G724S to be an Ex19Del-specific resistance mechanism to osimertinib. Retrospectively, we identified multiple patient cases now observed in the literature where patients with EGFR Ex19Del–mutant NSCLC displayed tumor progression post-osimertinib treatment in the presence of G724S (Piotrowska et al., 2018; Oztan et al., 2017; Peled et al., 2017; Fassunke et al., 2018). Together with the data we have presented here, these case studies suggest G724S functions as a resistance mutation in an allele-specific manner. To our knowledge, ours is the first evidence directly demonstrating that the underlying activating mutation (e.g. Ex19Del vs. L858R) influences the emergence of resistance mutations under selective pressure from a specific TKI.

Enhanced αC-helix stabilization in L858R results from polar interactions between the substituted arginine and neighboring negatively charged amino acids. In contrast, enhanced αC-helix stabilization in Ex19Del mutations likely results from alterations at the β3-αC interface. Structural superimposition of our active state deletion models onto EGFR WT shows that the position of L747 in WT is occupied by S752 (WT numbering) in E746_A750del and by the inserted valine in E746_S752>V (Supplementary Figure S8). Our data suggest that the P-loop conformational changes induced by G724S lead to destabilization of the αC-helix inward conformation in the presence of polar β3-αC interface substitutions.

Interestingly, Ex19Del/G724S displays phospho-EGFR levels similar to Ex19Del, but reduced phospho-EGFR compared to Ex19Del/C797S (Figure 2A). Our modeling suggests that stabilization of the αC-helix can vary between mutants upon introduction of G724S (Figure 2.11). Similarly, C797S may preferentially stabilize the αC-helix inward conformation of specific Ex19Del variants. C797 is a critical member of the structurally distinct catalytic spine (C-spine). The C-spine does not contribute to the interface formed by the glycine rich loop and β3-αC linker region. Nevertheless, previous network analysis by McClendon et al. (McClendon et al., 2014) suggests that the dynamics of the glycine rich loop and the C-spine may be highly correlated. We therefore suspect C797S may influence inter-domain correlations.

Fundamentally, our observations are similar to a concept familiar to clinical oncologists – that sequence variations in mutant proteins can impact drug binding. Osimertinib was developed to bind T790M with higher affinity than non-T790M EGFR mutants (Cross et al., 2014). Here, we show that sequence variations corresponding to the original activating mutation should also be taken into account when considering mechanisms of TKI resistance. Our findings have several important and immediate clinical implications. First, we further knowledge on a novel osimertinib resistance mutation that was not predicted by in vitro studies (Yu et al., 2007; Ercan et al., 2015). Recent studies have shown that G724S may be as prevalent as C797S in osimertinib resistant tumors (Fassunke et al., 2018). However, there are critical differences. While C797S containing EGFR mutants (e.g., Ex19Del/C797S) regain sensitivity to first-generation EGFR TKIs, erlotinib and gefitinib, the same G724S containing EGFR variants are cross-resistant to these inhibitors. In fact, there is an ongoing phase I clinical trial (NCT03122717) of osimertinib plus gefitinib combination therapy in patients with treatment naïve advanced EGFR-mutant NSCLC. This trial aims to test the hypothesis that circumventing C797S-mediated osimertinib resistance with gefitinib will prolong response. This concept will clearly not apply for patients with G724S mediated osimertinib resistance. However, our results support a role for afatinib therapy in treating Ex19Del patients with disease progression on osimertinib via C797S or G724S in the absence of T790M (Figure 2.5). Furthermore, in cases where G724S is potentially an independent oncogenic driver of other cancers, our results suggest possible treatment strategies with existing FDA-approved inhibitors. This level of evidence is critical to nominate variants of uncertain clinical significance, such as isolated G724S mutation, for eligibility into clinical trials such as NCI MATCH (NCT02465060).

These clinical consequences are rooted in structural perturbations to EGFR kinase. Detailed mechanistic understanding of these perturbations can provide critical insight to guide therapeutic intervention. Just prior to submission of the present manuscript, Fassunke et al. published investigations into the structural basis of EGFR G724S-mediated osimertinib (Fassunke et al., 2018). The authors coupled structure-based alignment of EGFR WT to EGFR D770_N771insNPG (exon 20 mutation) with P-loop RMSF calculations derived from short, single-trajectory cMD simulations. Specifically, Fassunke et al. demonstrated an elevated RMSF in both WT and E746_A750del when G724S is introduced. From that result, the authors postulated two potential, opposing mechanisms of G724S-mediated third-generation TKI resistance: (1) steric repulsion of the inhibitor, or (2) loss of important interactions with the inhibitor. However, RMSF calculations alone are rarely sufficient to provide detailed mechanistic insights (Farmer et al., 2017). Moreover, osimertinib resistance occurs in Ex19Del/G724S variants (Figure 2.5) but not G724S single mutants (Figure 2.10). The broad mechanisms previously posited do not provide adequate detail to address these data.

Here, we performed multiple independent GaMD enhanced sampling simulations in the presence and absence of osimertinib or afatinib totaling over 23 μs. For each EGFR mutant, we computed the relative binding

free energies of osimertinib and afatinib as well as the conformational free energy landscape profiles of the apo-state structures. While our RMSF calculations are consistent with Fassunke et al., our results further suggest that G724S hyper-stabilizes a β -bend conformation of the glycine-rich P-loop. This prevents contact of the F723 phenyl ring with the osimertinib indole ring. Our calculations suggest that L858R reversibly binds osimertinib with higher affinity than Ex19Del, and consequently loss of the F723 – osimertinib contact fails to disrupt binding in L858R. In Ex19Del, the addition of G724S destabilizes the reversible complex necessary for covalent adduct formation (Figure 2.2).

Moreover, we identified differences in P-loop conformational preferences between Ex19Del/G724S and L858R/G724S. (Figure 2.10A – D). In addition to our findings in Figure 2.2, it is possible that L858R/G724S is less poised to accommodate substrate binding vs. Ex19Del/G724S, resulting in L858R/G724S functioning as a catalytically inefficient receiver kinase in an asymmetric dimer; however, additional experiments would be required to test this hypothesis. It is also possible that L858R/G724S conformations may be less primed to support dimerization compared with L858R. The αC-helix of L858R/G724S bows outward over the course of the simulation, suggesting increased local instability. Despite still favoring the active state relative to WT, it is possible that with longer simulation times the αC-helix of L858R/G724S would more rapidly transition to a state incapable of supporting asymmetric dimerization than L858R (Figure 2.10C).

Importantly, our simulations also suggest that G724S increases the stability of the EGFR active conformation in the E746_S752 variant of Ex19Del, but reduces stability of the E746_A750del variant. Greater stability of the active αC-inward conformation in E746_S752>V/G724S offers a possible explanation for the enrichment of the rare variant Ex19Del in the Foundation Medicine cohort of NSCLC patients with G724S. Interestingly, of the four patients with genomic profiling data presented in Fassunke et al., all of them saw an increase in molecular fraction of G724S post-osimertinib therapy, and all of them had uncommon variants of Ex19Del (Fassunke et al., 2018).

These findings have implications in other, non-EGFR-mutant cancers as well. For example, ALK (anaplastic lymphoma kinase) rearrangements can be found in approximately 5% of NSCLC cancers (Lin et al., 2017). Over a dozen fusion partners have been identified across ALK+ cancers (Lin et al., 2017). Even the most frequently occurring fusion partner in ALK+ NSCLC, echinoderm microtubule-associated protein-like 4 (EML4), has $>$ 10 identified unique fusion variants (Shaw and Engelman, 2013). In addition, on-target acquired resistance to first- and second- generation ALK TKIs occurs in the form of approximately a dozen unique missense mutations (Gainor et al., 2016). Recent data suggests that a particularly recalcitrant ALK solvent front mutation, G1202R, is more likely to cause resistance in the context of EML4-ALK E6;A20 (V3) fusion rather than the more common EML4-ALK E13;A20 (V1) fusion (Lin et al., 2018). A structural basis for this observation was not presented; however, analogous to our current study, it could be that the unique

structural and biochemical properties of the original activating mutation foreshadowed the development of a specific resistance mutation.

In summary, we have employed an interdisciplinary computational and experimental approach which provides evidence that on-target osimertinib resistance in EGFR-mutant NSCLC occurs in an allele-specific manner dependent on the underlying activating mutation. Our data support a potential structural mechanism for Ex19Del/G724S osimertinib resistance, and open the door for further studies on TKI-EGFR interactions. We hope these mechanistic studies will be exploited to develop novel EGFR TKIs that circumvent multiple drug resistance mutations. Finally, we hope that insights from our investigations will be applied to develop increasingly effective targeted therapies for additional genetically-defined cancers.

2.4 Methods

2.4.1 Inhibitor source and preparation

EGFR TKIs were purchased from Selleck Chemicals (Houston, TX, USA). All drugs were prepared and stored as a stock solution at 10 mM in DMSO (Sigma-Aldrich, St. Louis, MO, USA).

2.4.2 Cell culture

293FT cells were purchased from Invitrogen (Carlsbad, CA, USA). NR-6 cells were a gift from Dr. William Pao (39). 293FT and NR-6 cells were cultured in DMEM with 4.5 g/L glucose, L-glutamine & sodium pyruvate (Mediatech, Corning, NY, USA) supplemented with 10% heat-inactivated fetal bovine serum (FBS) (Atlanta Biologicals, Flowery Branch, GA, USA) and penicillin (100 U/mL)/streptomycin (100 μg/mL) (Mediatech). Ba/F3 cells were purchased from DSMZ and were cultured in RPMI 1640 with L-glutamine (Mediatech) supplemented with 10% heat-inactivated FBS, penicillin (100 U/mL)/streptomycin (100 μg/mL), and 1 ng/mL interleukin-3 (IL-3) (Thermo Fisher Scientific, Waltham, MA, USA) until retroviral transduction and subsequent IL-3 withdrawal. Cells were grown in a humidified incubator with 5% CO2 at 37°C and were routinely evaluated for mycoplasma using a Venor GeM Mycoplasma Detection Kit (Sigma-Aldrich).

2.4.3 Immunoblot analysis

Cells were washed with PBS and lysed in radioimmunoprecipitation analysis buffer (50 mM TrisHCl pH 8.0, 150 mM sodium chloride, 5 mM magnesium chloride, 1% Triton X-100. 0.5% sodium deoxycholate, 0.1% SDS, 40 mM sodium fluoride, 1 mM sodium orthovanadate, and complete Protease Inhibitor Cocktail [Roche Diagnostics, Indianapolis, IN, USA]). Western Lightning ECL reagent (PerkinElmer, Waltham, MA, USA) was used for signal detection. β-actin antibody (A2066) was purchased from Sigma-Aldrich. EGFR (#2232), pEGFR Y1068 (#2234), pEGFR Y1173 (#2244), ERK (#9102), pERK T202/Y204 (#9101), horseradish

peroxidase (HRP)-conjugated anti-mouse (#7076) and HRP-conjugated anti-rabbit (#7074) antibodies were purchased from Cell Signaling (Danvers, MA, USA). Each experiment was performed twice.

2.4.4 CellTiter Blue cell viability assay

Ba/F3 cells were seeded in 96-well plates at a density of 20,000 cells/well and treated with varying concentrations of indicated compounds, with six technical replicates per concentration. After 72 hours, CellTiter Blue Reagent (Promega, Madison, WI, USA) was added to wells according to manufacturer's instructions, and cells were incubated at 37°C with 5% CO2 for 2 to 4 hours. Absorbance was detected at 590 nm with a Synergy HTX microplate reader (BioTek Instruments, Winooski, VT, USA). Each experiment was performed three times.

2.4.5 Statistical analysis

All experiments were performed at least three times and the differences were determined by one-way ANOVA. Differences were considered significant when $p < 0.05$.

2.4.6 Molecular Modeling

Structural models of the EGFR kinase exon 19 deletion mutants (Ex19Del) were generated through complementary use of the structure-prediction software package Rosetta utilizing the REF2015 score function (40-42) and molecular dynamics (MD) simulation with AMBER16 (43). Comparative models of Ex19Del kinase domain were created with RosettaCM (40,41) by modeling the kinase domain sequence sans β3-αC residues E746–A750 for the canonical variant model, or a valine substituted for the range E746–S752 for the rare variant model, and applying PDB IDs 2GS6 and 2GS7 as templates for the active and inactive state models, respectively (31). Active and inactive state Rosetta models of EGFR were minimized and allowed to equilibrate in a rectangular box of TIP4PEW explicit solvent neutralized with monovalent chlorine anions (44,45). Solute was buffered on all sides with 12 Å solvent. Afterward, dual-boost Gaussian accelerated MD (GaMD) simulations were performed to enhance conformational sampling (23,24,46,47). Protein-ligand binding free energy calculations were performed with MM/GBSA implemented in the AMBER suite in combination with the quasi-harmonic approximation (QHA) of entropy (48). For a detailed description of model building, molecular dynamics simulations, and binding free energy calculations, please see the Supplementary Methods section.

2.4.7 Genomic profiling of patient samples

Hybrid capture-based next generation sequencing (NGS) was performed on formalin-fixed paraffin embedded tissue sections or circulating tumor DNA isolated from blood samples in a Clinical Laboratory Improvement

Amendments (CLIA)- certified, CAP (College of American Pathologists)-accredited laboratory (Foundation Medicine, Cambridge, MA) as described previously (49,50). Approval for this study, including a waiver of informed consent and a HIPAA waiver of authorization, was obtained from the Western Institutional Review Board (Protocol No. 20152817).

CHAPTER 3

Allele-specific activation and inhibitor sensitivities of EGFR exon 19 deletion mutations in lung cancer

This chapter is a collaborative work of Benjamin P. Brown*, Yun-Kai Zhang*, Soyeon Kim*, Patrick Finneran, Yingjun Yan, Zhenfang Du, Jiyoon Kim, Abigail Leigh Hartzler, Michele L. LeNoue-Newton, Adam W. Smith, Jens Meiler, and Christine M. Lovly (*These authors contributed equally).

3.1 Introduction

Epidermal growth factor receptor (EGFR) mutations are responsible for 15 – 30% of all cases of non-small-cell lung cancer (NSCLC) (Pao et al., 2004; Lynch et al., 2004). Of these mutations, >90% can be attributed to either the L858R mutation in the kinase domain (KD) activation loop (A-loop), or deletion/insertion mutations in exon 19 (henceforward categorically referred to as ex19del mutations) corresponding structurally to the β3-αC loop in the KD (Pao et al., 2004; Lynch et al., 2004). Historically, ex19del mutations have not been clinically differentiated. In the first clinical trials to establish the superior efficacy of EGFR tyrosine kinase inhibitors (TKIs) compared to chemotherapy, EGFR KD oncogenic mutations were all considered interchangeable (Mitsudomi et al., 2010). Today, the current clinical standard of care for EGFR-mediated NSCLC is osimertinib. The seminal phase 3 clinical trial that demonstrated osimertinib's increased efficacy compared to standard gefitinib or erlotinib TKI therapy, FLAURA, did separately annotate and compare L858R and ex19del (Soria et al., 2018; Ramalingam et al., 2020); however, heterogeneity within the ex19del group was not considered.

This is in stark contrast to the less frequently occurring EGFR exon 20 insertion (ex20ins) mutations. It has been appreciated in the literature that ex20ins display heterogeneity in enzyme activity, clinical phenotype, and sensitivity to existing FDA-approved TKIs (He et al., 2012; Kosaka et al., 2017; Naidoo et al., 2015; Yasuda et al., 2012, 2013). At the structural level, molecular dynamics (MD) simulations suggest that ex20ins mutants can lower the free energy barrier associated with adopting the KD active conformation in an allele-specific manner (Ruan and Kannan, 2018). There are multiple ongoing drug development efforts aimed at designing TKIs to treat ex20ins-mediated cancers differently (Gonzalvez et al., 2021; Riely et al., 2021; Jang et al., 2018). Several retrospective studies have now suggested that there are differences in patient outcomes between ex19del patient populations (Tokudome et al., 2020; Zhao et al., 2020; Xu et al., 2020; Chung et al., 2012; Su et al., 2017; Stewart et al., 2018). Not surprisingly then, emerging evidence suggests that the lack of allele-specific resolution of ex19del variants in clinical practice can impede our ability to provide optimal therapeutic strategies for NSCLC and other cancer patients.

It is also noteworthy that investigations into ex19del often use the verbiage "exon 19 deletion" to refer to different allele variants, making it more challenging to functionally characterize them and develop appropriate therapeutic strategies. For example, the mechanism of activation of ex19del has been reported to be both ligand-independent (Cho et al., 2013; Greulich et al., 2005; Valley et al., 2015; Okabe et al., 2007) and ligand-dependent (Sordella et al., 2004; Carey et al., 2006; Mulloy et al., 2007), and it is unclear to what extent the discrepancy is a result of the use of different experimental methodologies or different ex19del variants. We have also previously found that the development of osimertinib resistance to the G724S mutant is dependent on the specific ex19del variant (Brown et al., 2019a), suggesting that ex19del structural differences can have therapeutic implications. Thus, to maximize the efficacy of targeted therapies we need to refine our understanding of oncogenic variants at the atomic level.

In this study, we tested the hypothesis that sequence variation between EGFR oncogenic ex19del mutations can lead to allele-specific activation and TKI sensitivity. We probed the AACR GENIE database (31) and identified 60 unique ex19dels and built structural models of each variant. Next, we selected three of the most common variants predicted to be structurally distinct for detailed computational, biophysical, and biochemical evaluation: E746_A750, E746_S752>V, and L747_A750>P. Altogether, our results demonstrate that ex19dels are a functionally heterogeneous population with potentially unique considerations for optimal therapeutic targeting.

3.2 Results

3.2.1 ex19del sequence variants cluster by chemical conservation and thus function

We first investigated the sequence heterogeneity of ex19del variants by probing the AACR GENIE database (Consortium, 2017). We identified 60 variants and mapped these variants to the EGFR kinase domain (KD) (Figure 3.1). Structurally, exon 19 corresponds to the β3 sheet, β3-αC loop, and N-terminal half of the αC helix (Figure 3.1A). All residues are numbered with respect to WT in the immature form (e.g., we reference L858R instead of L834R). We identified mutants ranging in size from a single residue deletion to a net eight residue deletion. The starting and stopping points for the deletions predominantly occurred at residues E746, L747, A750, T751, S752, and P753, such that the length of the β3-αC loop is highly subject to sequence variation in comparison to the β3 or αC regions (Figure 3.1B). The predominant mutations are E746_A750 (62.9%), L747_P753>S (7.4%), L747_T751 (5.2%), E746_S752¿V (4.0 %), and L747_A750>P (3.7%) (Figure 3.1C).

The breadth of variants is substantial, ranging from deletions that occur entirely in β3 (K739_I744>N) to those occurring almost entirely in αC (e.g., P753_I759). To help characterize the mutations, we first built structural models of all variants utilizing the Rosetta comparative modeling approach coupled with Gaussian

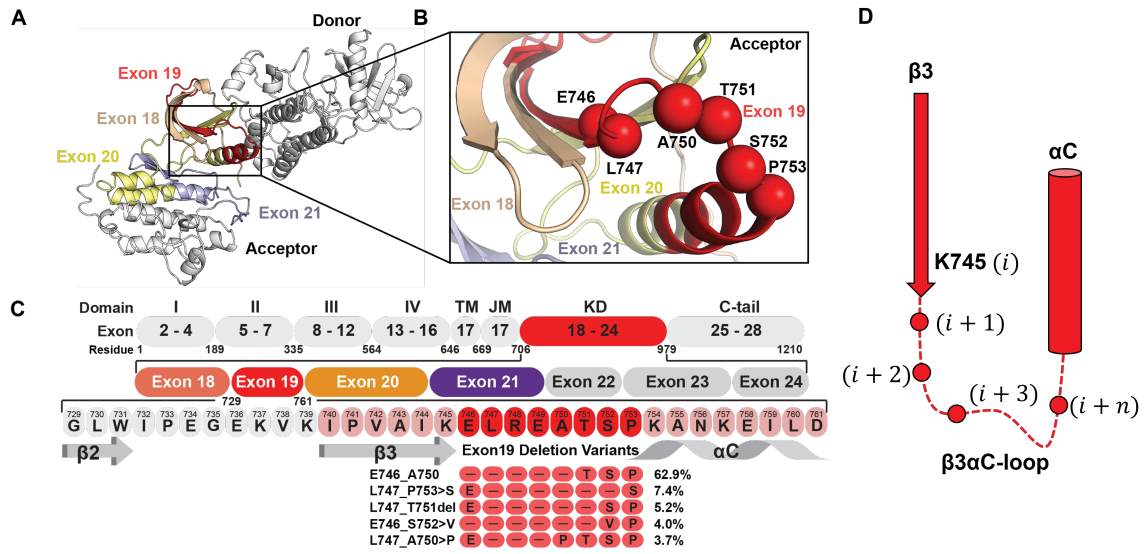


Figure 3.1: Frequently occurring mutations in the EGFR β3-αC motif. (A) Schematic representation of the active EGFR-WT asymmetric dimer. Oncogenic and TKI resistance mutations have been reported in exons 18 (wheat), 19 (red), 20 (yellow), and 21 (blue). (B) The majority of deletion mutations begin at residues E746, L747, or T751. Deletion mutants frequently terminate with or without an insertion at position A750, T751, S752, or P753. Spheres indicate the residue Cα. (C) Multiple sequence alignment of the β3-αC motif between EGFR-WT and ex19del variants with >2% frequency. (D) Residues at the β3αC interface can be referenced with respect to their index after the conserved K745 residue in the majority of mutants.

accelerated MD (GaMD) (Miao et al., 2015) (see Methods). Our models suggested several recurring structural features of ex19del. First, the most common ex19del variants, including E746_A750, L747_P753>S, and L747_T751 (Figure 3.1C), replace L747 at the β3-αC interface with a serine and simultaneously remove at least one full turn from the N-terminus of the αC helix (Figure 3.2A). Second, mutants with net deletions of size three, such as L747_A750>P and E746_T751>APS, frequently converge on the same β3-αC loop conformation, characterized by a β3-αC tight turn with proline in the second position (Figure 3.2B). Third, we observed that several mutants project polar residues into the ATP binding pocket in the vicinity of the canonical K745 – E754 salt bridge, such as L747_S752>Q and E746_S752>V (cis-trans proline-dependent).

To deeply evaluate potential functional differences between mutants, we selected three mutants that are prevalent in patients based on our AACR GENIE analysis (Figure 3.1C) and cover the breadth of features described above: E746_A750, E746_S752>V, and L747_A750>P. For clarity, we periodically reference residues by their position relative to K745 (Figure 3.1D).

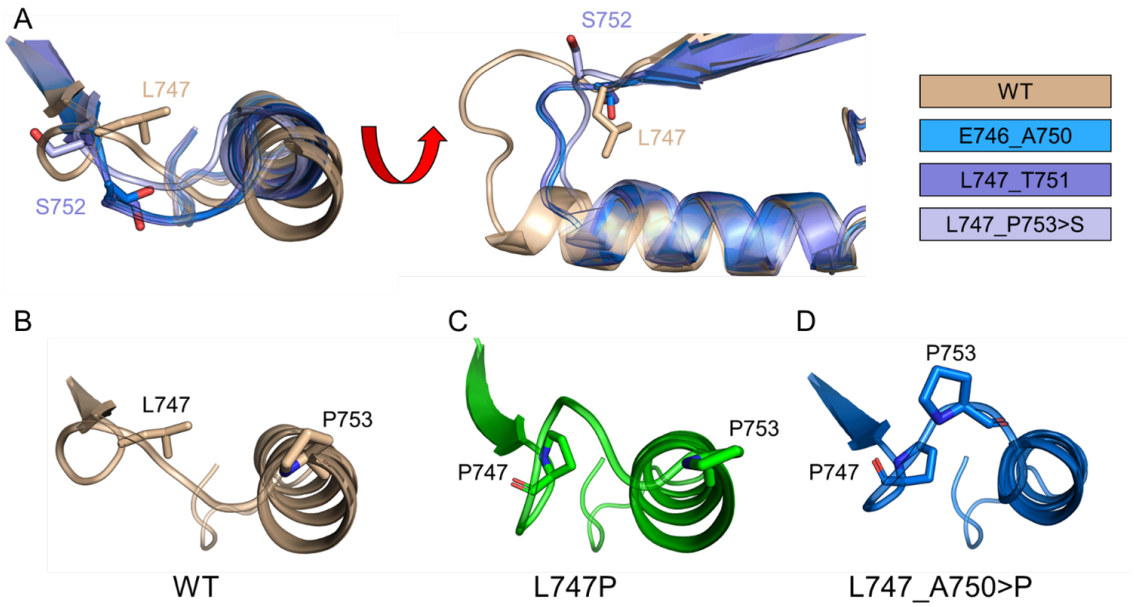


Figure 3.2: Structural comparison of modeled ex19del $\beta 3\alpha$C motifs. (A) Superimposition of the $\beta 3\alpha$C region of the most common ex19del variants with WT. Rendering of the $\beta 3\alpha$C loop in (B) WT, (C) L747P, and (D) L747_A750>P. L747P and L747_A750>P both form a tight turn in the $\beta 3\alpha$C loop. The L747_A750>P tight turn contains a proline in the second position and fewer residues on the N-terminus of the αC-helix.

3.2.2 ex19del variants adopt unique $\beta 3$-αC conformations with different energetic barriers to activation

We began with the hypothesis that ex19dels can display allele-specific differences in their propensity to adopt the active conformation. Wild-type EGFR (WT) is activated when ligand binds the extracellular domain (ECD) to promote intermolecular dimerization and multimer/oligomerization (Cohen S Fau Carpenter et al., 1980; Needham et al., 2016a; Huang et al., 2016). Intracellularly, this results in asymmetric dimerization between two KD where the "receiver" KD is stabilized in an active conformation by the "donor" KD (Zhang et al., 2006). Previous investigations have shown that oncogenic variants in the KD often stabilize the αC-helix by suppressing intrinsic disorder (Shan et al., 2012) leading to enhanced dimerization where the mutant KD behaves as a "super acceptor" (Red Brewer et al., 2013).

Subsequently, we performed six (E746_A750, E746_S752>V, and L747_A750>P, in active and inactive state respectively) independent conventional molecular dynamics (cMD) simulations of 4.0 – 6.0 s for each structure, such that three simulations were initiated from each state (120.0 s total). Consistent with previous reports (Shan et al., 2013), the αC helix of WT readily departed from the active conformation to adopt an unstructured intermediate state, and 1/3 active state simulations transitioned completely to the Src-like inactive conformation (αC helix out, A-loop in, DFG in) (Figure 3.3A).

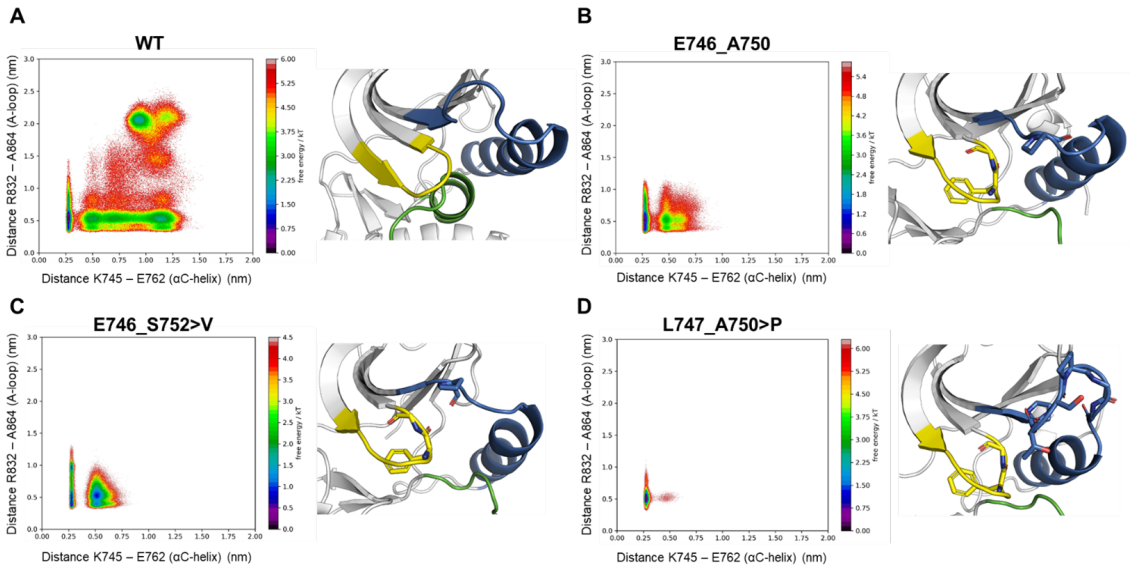


Figure 3.3: Conventional MD simulations of several ex19del variants starting from the active state. Boltzmann-weighted probability distributions of (A) WT, (B) E746_A750, (C) E746_S752>V, and (D) L747_A750>P conformational changes in conventional MD simulations. All simulations were started from the active state. Three independent simulations for each system were run for 4.0 us each. The inward/outward motion of the activation loop is depicted on the y-axis (larger numbers indicate more inward), and the inward/outward motion of the αC-helix is depicted on the x-axis (larger numbers indicate more outward). Snapshots are from the end of one of the three independent simulations. WT transitioned to the Src-like inactive state in one of the three simulations. The glycine-rich loop is colored yellow, the $\beta 3\alpha$C-loop and αC-helix are blue, and the activation loop is green.

In comparison, each of the ex19del variants remained stable in the active state (αC helix in, A-loop out, DFG in, Figure 3.3B – D). The tight turn predicted in the Rosetta/GaMD model of L747_A750>P is highly stable, preventing inactivation (Figure 3.3D). Unfortunately, no transitions were observed from the inactive to the active state or vice versa in any of the ex19del cMD simulations. Therefore, we combined steered MD (SMD) with umbrella sampling (UMD) simulations to map the conformational free energy landscape (FEL) of the transition.

Following a procedure similar to that previously employed for ex20ins variants (Ruan and Kannan, 2018) we defined our umbrella sampling collective variables (CV) along two dimensions: (1) Activation state of the αC helix as defined by the difference in distance between K860 – E762 and K745 – E762, and (2) activation state of the A-loop as defined by the dihedral angle formed by the Cα atoms of D855 – F856 – G857 – L858 (Figure 3.4A, B).

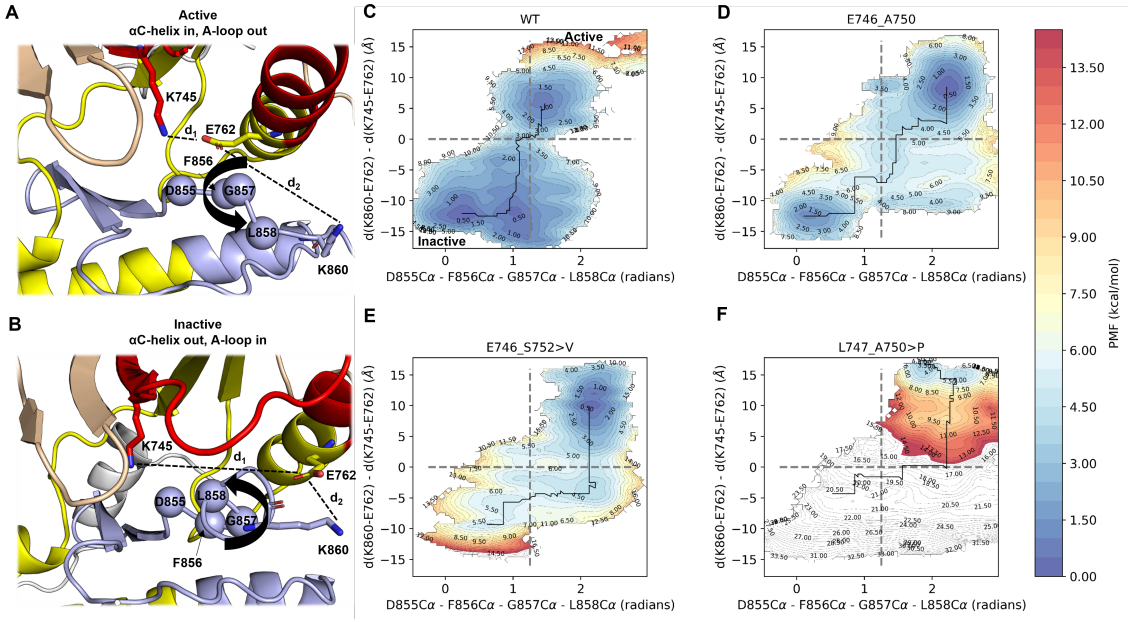


Figure 3.4: Conformational free energy landscapes of ex19del variants from umbrella sampling MD simulations. Collective variables describe the (A) active and (B) inactive states as the pseudo-dihedral angle formed by the alpha carbon atoms of residues D855, F856, G857, and L858 (x-axis) as well as the difference in distance between the capping sidechain atoms of E762 and K745 (d1) and E762 and K860 (d2) (y-axis). Conformational free energies are shown for (C) WT, (D) E746_A750, (E) E746_S753¿V, and (F) L747_A750>P. Plots are contoured at 0.5 kcal/mol and colored within the range 0 (blue) and 15 (red) kcal/-mol. Contours above 15 kcal/mol are colored white.

Using these 2 CVs, we measured the free energy difference between the active and inactive states of WT and found it to be approximately 1.0 kcal/mol in favor of the inactive state (Figure 3.4C), in good agreement with prior estimates (Ruan and Kannan, 2018). In contrast to WT and the previously reported exon 20 insertion mutations (Ruan and Kannan, 2018), all three ex19del variants favored the active state

(Figure 3.4D – F). E746_A750 and E746_S752>V favored the active state by approximately 1.0 kcal/mol and 4.5 kcal/mol, respectively (Figure 3.4D – E). We also performed SMD+UMD simulations on the other two most commonly occurring ex19dels, L747_P753>S and L747_T751. L747_T751 displays an activation profile similar to E746_S752>V, while L747_P753>S may be more comparable to several ex20ins variants (Ruan and Kannan, 2018) (Figure 3.5).

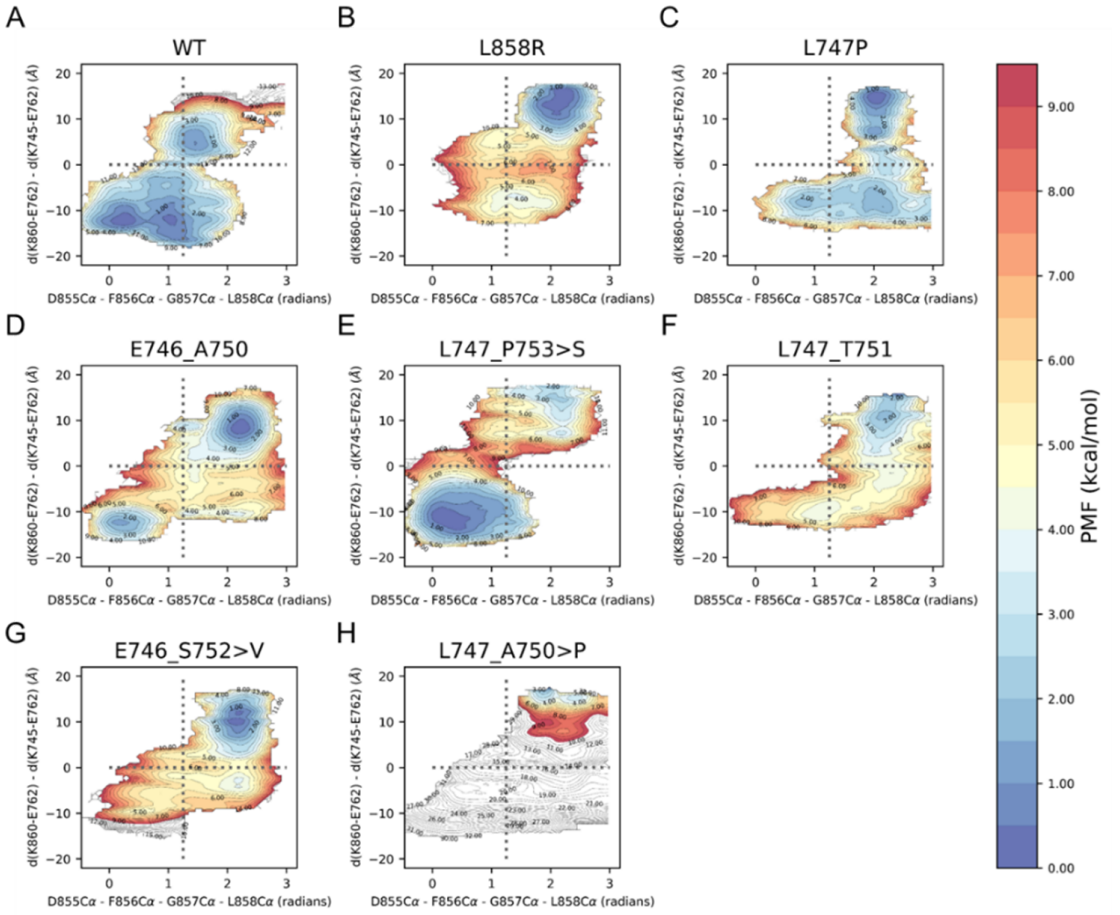


Figure 3.5: Conformational free energy landscapes of EGFR variants from umbrella sampling MD simulations. Collective variables describe the active and inactive states as the pseudo-dihedral angle formed by the alpha carbon atoms of residues D855, F856, G857, and L858 (x-axis) as well as the difference in distance between the capping sidechain atoms of E762 and K745 (d1) and E762 and K860 (d2) (y-axis). Conformational free energies are shown for (A) WT, (B) L858R, (C) L747P, (D) E746_A750, (E) L747_P753>S, (F) L747_T751, (G) E746_S752>V, and (H) L747_A750>P. Plots are contoured at 0.5 kcal/mol and colored within the range 0 (blue) and 9.5 (red) kcal/mol. Contours above 9.5 kcal/mol are colored white.

Interestingly, L747_A750>P appears to be trapped in the active state, with prohibitively large free energy barriers to the inactive state (Figure 3.4F). We considered that this may be a result of the proline substitution at position 747. We tested this hypothesis by building models for the oncogenic missense variant L747P (Liang et al., 2019) and performing SMD+UMD simulations. L747P induces an ordered tight turn in the β3-αC

loop, stabilizing the active state over the inactive state by approximately 1.0 kcal/mol (Figure 3.5C), but not by as large a margin as L747_A750>P. The substantially larger barrier to inactivation in L747_A750>P may result from the proline in its $\beta 3\alpha$C tight turn coupled with the net three residue deletion (Figure 3.2B). Altogether, our results suggest that ex19del variants adopt unique conformations near the receiver KD interface that translate into potentially substantial differences in activation propensity.

3.2.3 L747_A750>P, but not E746_A750 or E746_S752>V, dimerizes in a ligand-independent manner

Previous studies have suggested that KD mutants may promote ligand-dependent "inside-out" dimerization (Tsai and Nussinov, 2019). Based on our simulation results, we hypothesized that the L747_A750>P variant forms dimers in the absence of ligand stimulation because it is trapped in a receiver kinase active state. To test our hypothesis, we measured the homo-interaction stoichiometry of each variant in the presence and absence of EGF ligand using two-color pulsed interleaved excitation fluorescence cross-correlation spectroscopy (PIE-FCCS) (Huang et al., 2016; Du et al., 2021). Live cell PIE-FCCS measurements and analysis were completed on single cells expressing individual ex19del variants with WT data recorded as a negative control for each experiment (see Methods).

First, we performed PIE-FCCS experiments in the absence of EGF ligand. Samples were serum starved for 24 hours to ensure no residual ligand-dependent effects. As expected, WT has a median cross-correlation (fc) value near zero (fc = 0.01), indicating that it exists predominantly as a monomer. Our results also suggest that E746_A750 and E746_S752>V are predominantly monomeric in the absence of ligand (fc = 0.05 and 0.06, respectively). In contrast, L747_A750>P displays significantly higher median cross-correlation (fc = 0.13) (Figure 3A.6). Consistent with the cross-correlation values, the diffusion coefficients of eGFP-tagged WT (0.35 $\mu m^2/s$), E746_A750 (0.35 $\mu m^2/s$), and E746_S752>V (0.33 $\mu m^2/s$) are significantly higher than L747_A750>P (0.18 $\mu m^2/s$) (Figure 3B.6). The increased median cross correlation and decreased diffusion coefficient of L747_A750>P relative to WT is indicative of dimer formation in the absence of ligand stimulation.

Next, we performed PIE-FCCS experiments in the presence of EGF ligand to evaluate whether or not ex19del variants differ in their response to extracellular stimulation. A recent study demonstrated that KD mutations can directly change the conformational preferences of the ECD, potentially modulating signaling responses to ligand (Huang et al., 2020). Here, we observed that WT forms multimers upon stimulation with EGF, consistent with prior studies (fc = 0.31; D = 0.13 $\mu m^2/s$) (Needham et al., 2016a; Huang et al., 2016; Du et al., 2021; Clayton et al., 2005). EGF stimulation caused E746_A750 (fc = 0.16; D = 0.23 $\mu m^2/s$) E746_S752>V (fc = 0.17; D = 0.18 $\mu m^2/s$), and L747_A750>P (fc = 0.18; D = 0.17 $\mu m^2/s$) to form a mixture of dimers and multimers (Figure 3A.6, B). The fact that each of the mutants show lower cross-

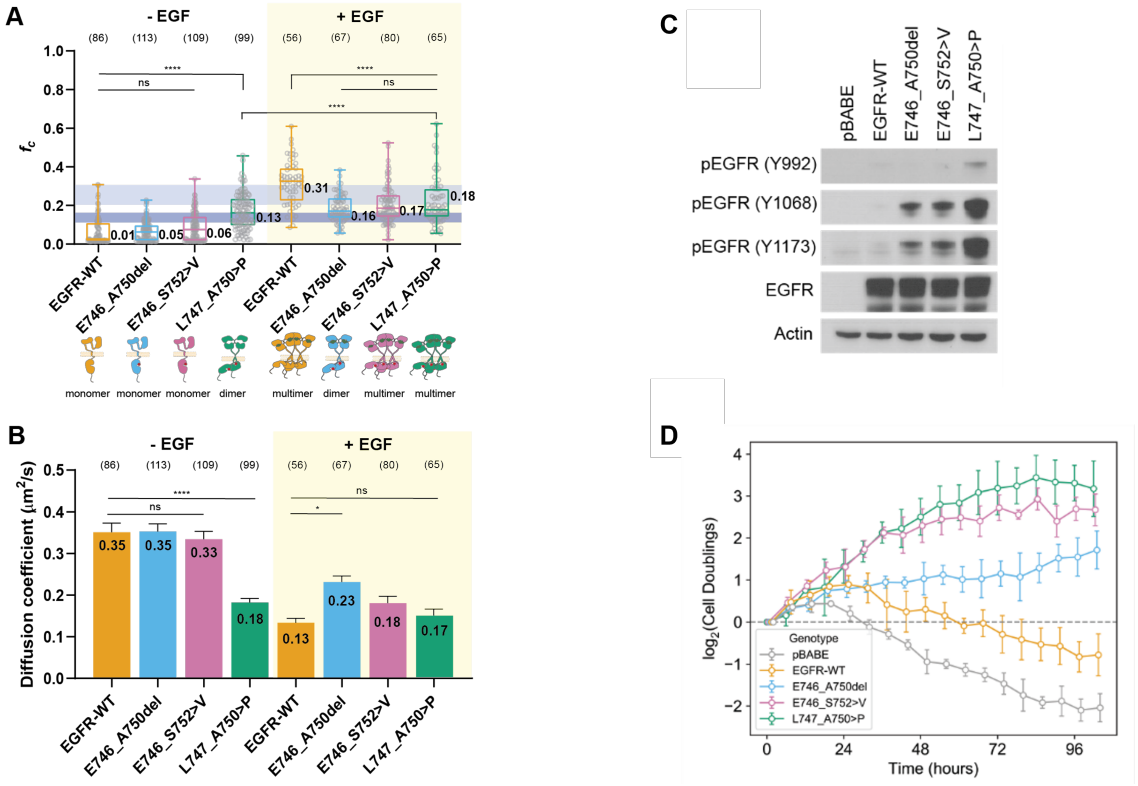


Figure 3.6: Ex19del variants display allele-specific differences in dimerization and oncogenic growth. (A) Cross correlation values of transfected EGFR variants with (+) or without (-) ligand (EGF) stimulation. The dark and light blue boxes indicate the fc value regions for dimers and multimers, respectively. (B) Diffusion coefficient values of EGFR variants with (+) or without (-) ligand (EGF) stimulation. The light orange box indicates EGF-stimulated groups. (C) Ba/F3 cells were stably transfected with different EGFR ex19del variants, WT, or empty vector. Cellular lysates were probed with the indicated antibodies to measure phosphorylation. (D) Rate of IL-3-independent growth of Ba/F3 cells stably transfected with different ex19del variants, WT, or empty vector. Data and illustrations for figure panels A and B produced by Soyeon Kim, Abigail Leigh Hartzler, and Adam W. Smith. Data and illustrations for figure panels C and D produced by Yun-Kai Zhang, Yingjun Yan, Zhenfang Du, Jiyoon Kim, and Christine M. Lovly.

correlation and faster diffusion compared to WT suggests that the ex19del mutations have a significant effect on the formation of ligand-depended multimeric assemblies.

3.2.4 E746_S752>V and L747_A750>P display enhanced oncogenic activation relative to E746_A750

The strong energetic preference of L747_A750>P to adopt the active conformation (Figure 3.4E) and corresponding propensity to form ligand-independent dimers (Figure 3A.6) led us to hypothesize that L747_A750>P would display enhanced oncogenic growth compared with other ex19del variants in vitro. To test our hypothesis, we generated expression vectors containing empty vector, WT, E746_A750, E746_S752>V, or L747_A750>P and introduced these into murine lymphoid Ba/F3 cells (45). After selection of stable expression in puromycin, the cells were collected, lysed and blotted for EGFR autophosphorylation (pEGFR). Our results confirmed that all three ex19del variants exhibit strong pEGFR compared to WT. In support of our hypothesis, we observed that L747_A750>P displays substantially higher levels of pEGFR compared with either E746_A750 or E746_S752>V (Figure 3C.6).

To further investigate ex19del variant differences in IL-3 independent oncogenic growth in Ba/F3 cells, we depleted IL-3 from the growth medium to monitor changes in cell counts over time (Figure 3D.6). As expected, the Ba/F3 cells expressing either vector or WT EGFR died shortly upon withdrawal of exogenous IL-3, while cells expressing EGFR ex19del variants survived and proliferated. Cells expressing either E746_S752>V or L747_A750>P proliferated at a higher rate compared with cells expressing E746_A750del (Figure 3D.6). Despite not undergoing ligand-independent dimerization as did L747_A750>P in PIE-FCCS experiments, cells expressing E746_S752>V displayed statistically similar growth rates compared with L747_A750>P. Collectively with our MD simulations, our results suggest that ex19del variants differentially promote growth and enzymatic activity as a function of their energetic barriers to activation.

3.2.5 E746_S752>V and L747_A750>P are less sensitive to TKI treatment than E746_A750

We considered the possibility that differences may also exist between ex19del variant TKI sensitivities, which also may explain differences in outcomes between patients with specific ex19dels (17, 21). We previously found that some ex19del variants, in particular E746_S752>V, are especially likely to develop G724S-mediated resistance in response to osimertinib, while L858R and other ex19del variants are not (Brown et al., 2019a; Fassunke et al., 2018). Recently, it was further suggested that L747_A750>P has reduced sensitivity to erlotinib and osimertinib relative to E746_A750 in functional assays due to steric effects (Truini et al., 2019). Thus, we sought to evaluate the relative TKI sensitivity of E746_A750 in comparison to E746_S752>V and L747_A750>P.

We first treated Ba/F3 cells expressing E746_A750, E746_S752>V, or L747_A750>P with either 30 or

100 nM osimertinib. We observed that autophosphorylation was markedly reduced in both E746_A750 and L747_A750>P, but not in E746_S752>V (Figure 3.7A). Subsequently, we performed the same experiment in well-established lung adenocarcinoma cell lines expressing E746_A750 (PC9), E746_S752>V (SH450), or L747_A750>P (HCC4006). Again, we observed that E746_S752>V was less sensitive to osimertinib than E746_A750 or L747_A750>P. To model the clinical exposure of EGFR TKIs in lung adenocarcinoma, we performed long-term treatments of osimertinib in these cell lines at a clinically relevant dose (100 nM) (48) with periodic medium/TKI refreshment (Figure 3.7C). The untreated PC9, SH450, and HCC4006 cells underwent exponential growth and quickly reached confluence within 3 days. The growths of PC9 and HCC4006 cells were inhibited effectively by osimertinib treatment, and the cells initially stopped growing. In particular, the proliferation of PC9 cells was successfully inhibited by osimertinib for more than three weeks. We observed that the HCC4006 cells gradually adapted to the treatment and proliferated to confluence in 20 days. Most notably, however, osimertinib only partially inhibited the proliferation of SH450 cells, and after an incomplete response continued growing, reaching confluence within a week. Thus, consistent with our Western blots, we found that E746_S752>V was least responsive to osimertinib, followed by L747_A750>P, while E746_A750 was completely inhibited (Figure 3.7C).

Based on our in vitro data, we hypothesized that E746_S752>V has a lower osimertinib binding affinity than E746_A750 and L747_A750>P. To test this hypothesis, we performed MD simulations of each of the ex19del variants in complex with osimertinib. We performed three independent MD simulations of 2.0 s each for each EGFR variant (WT, E746_A750, E746_S752>V, E746_S752>V/G724S, or L747_A750>P) bound to osimertinib starting from either the active or inactive conformation (sans inactive E746_S752>V/G724S; 60.0 s aggregate simulation time). As expected based on the available crystallographic evidence (Yosaatmadja et al., 2015), osimertinib binding energies were estimated to be better in the active state than the inactive state in all cases. Both E746_A750 and L747_A750>P were estimated to have a better osimertinib binding free energy than WT (Figure 3.7E). Contrary to our hypothesis, E746_S752>V was not predicted to bind osimertinib with a lower affinity than E746_A750 or L747_A750>P. In contrast to previous studies (Truini et al., 2019), L747_A750>P was not estimated to have reduced osimertinib binding free energy (Figure 3.7E).

To better understand our simulation results, we quantitatively evaluated the inhibitory efficacy of three generations of EGFR TKIs (erlotinib, afatinib, and osimertinib) by measuring cell viabilities of isogenic Ba/F3 cells stably transfected with either E746_A750, E746_S752>V, or L747_A750>P in the presence of each TKI separately. We observed that L747_A750>P and E746_S752>V were both at least 10x less sensitive to TKI than E746_A750 (Figure SX). We corroborated these results by measuring cell viabilities of lung adenocarcinoma cell lines expressing different ex19del variants. Here, we also observed that SH450 (E746_S752>V) or HCC4006 (L747_A750>P) were at least 10x less sensitive to erlotinib than PC9

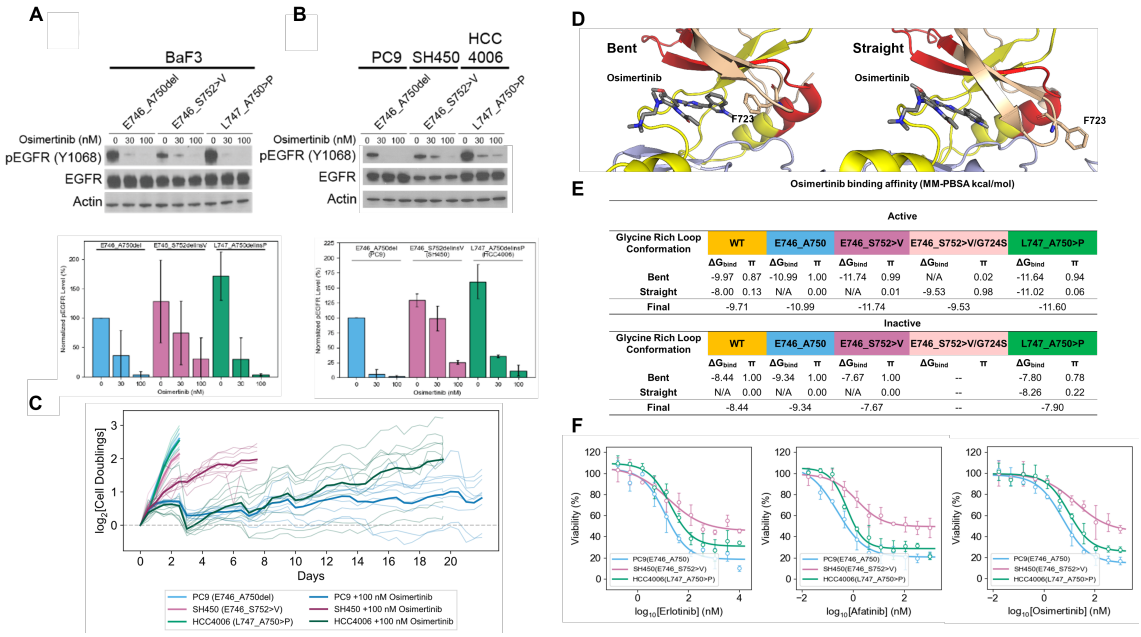


Figure 3.7: Allele-specific differences in ex19del TKI sensitivity may not be due to differences in TKI binding affinity. (A) Ba/F3 cells were stably transfected with different EGFR ex19del variants and treated with increasing concentrations (0, 30, or 100 nM) of osimertinib. Cellular lysates were probed with the indicated antibodies to measure phosphorylation. (B) Lung adenocarcinoma cell lines expressing E746_A750 (PC9), E746_S752>V (SH450), or L747_A750>P (HCC4006) were treated with increasing concentrations (0, 30, or 100 nM) of osimertinib. Cellular lysates were probed with the indicated antibodies to measure phosphorylation. Quantifications are represented as the average grayscale ratio of pEGFR/EGFR/Actin+/- standard deviation across three independent biological replicates. (C) Time-dependent growth of lung adenocarcinoma cell lines expressing E746_A750 (PC9), E746_S752>V (SH450), or L747_A750>P (HCC4006) treated with either 100 nM osimertinib or buffer. Each condition was performed with 9 replicates (thin lines) and averaged (bold lines). (D) Structural models of EGFR in complex with osimertinib in either the bent (F723 facing osimertinib in the ATP binding pocket) or straight (F723 projecting away from the ATP binding pocket) conformations. (E) Osimertinib binding affinities for each ex19del variant, WT, and the double mutant E746_S752>V/G724S from simulations starting in the active and inactive states. Bent and straight states were separated by a small 2-state Markov state model based on the G/S724 backbone phi angle. MM-PBSA was not performed if the stationary distribution for a state was estimated at less than 0.05 or the model failed to pass a Chapman-Kalmogorov test. Binding energies are computed as the average MM-PBSA energies of 1000 randomly selected frames from the corresponding MSM cluster. For each EGFR variant, six simulations of 2.0 us each were performed such that there were three each from the active and inactive states (except E746_S752>V/G724S, for which no inactive state simulations were performed). (F) Cell viability assays performed in lung adenocarcinoma cell lines stably expressing E746_A750 (PC9), E746_S752>V (SH450), or L747_A750>P (HCC4006) with first (erlotinib), second (afatinib), and third (osimertinib) generation EGFR TKIs. Data and illustrations for figure panels A, B, C, and F produced by Yun-Kai Zhang, Yingjun Yan, Zhenfang Du, Jiyoon Kim, and Christine M. Lovly.

(E746_A750). SH450 were also greater than 10x less sensitive to afatinib and osimertinib as compared to PC9 or HCC4006 (Figure 3.7F). L747_A750>P displays a similar response to afatinib as E746_A750. Our results suggest that E746_S752>V and L747_A750>P are intrinsically less sensitive to ATP-competitive TKIs in vitro. E746_A750 displays the most TKI sensitivity among the three ex19dels.

3.2.6 Differences in ATP binding may modulate TKI sensitivity across ex19del variants

Our in vitro data suggest that E746_S752>V and L747_A750>P display intrinsic resistance to standard first-, second-, and third-generation TKIs. Simultaneously, our MD simulations estimate that E746_S752>V and L747_A750>P reversibly bind osimertinib at least as well as E746_A750. Thus, we hypothesized that the reduced sensitivity of E746_S752>V or L747_A750>P to ATP-competitive inhibitors is the result of higher ATP binding affinities in these receptors than other EGFR oncogenic variants, thereby reducing the relative binding affinity of TKI to ATP.

To test this hypothesis, we estimated the apparent ATP Km and erlotinib Ki for WT, E746_A750, L747_A750>P, and an additional uncommon variant L747_E749 using the ADP-Glo assay as described in the Methods. We chose erlotinib for the TKI binding affinity analysis to enable explicit comparison of the effects of ATP Km on noncovalent TKI interactions. Our ADP-Glo assay results suggest that there are substantial differences in ATP kinetics between EGFR variants, consistent previous reports on L858R and G719S (Carey et al., 2006; Yun et al., 2008a).

E746_A750 and L747_E749 display ATP Km values of 100 μM. In contrast, L747_A750>P displays an ATP Km of 6 μM. Interestingly, the rate of phosphate transfer in L747_A750>P is 17x lower than E746_A750, but the reduced Km results in comparable catalytic efficiencies (Table 3.1). In contrast to ATP Km, the difference in erlotinib binding is comparatively small between the tested variants (all within a factor of 2 to one another). This results in the apparent erlotinib potency, taken as the ratio of Ki to ATP Km, to be 18x lower in L747_A750>P than E746_A750 (Table 3.1). These data are consistent with the reduced sensitivity of L747_A750>P in vitro and suggest a general mechanism by which ex19del variants may differ in their responses to TKI.

EGFR	Amount	ATP Km (μM)	ATP Vmax	ATP Vmax/Km	ERL Ki (nM)	Ki/(Km*10^3)	Ki/Km Norm E746_A750
WT	25ng/rxn	54.05	4.63E-02	8.57E-04	6.12	1.13E-04	2.57
E746_A750	25ng/rxn	105.30	3.76E-02	3.57E-04	4.63	4.40E-05	1.00
L747_A750>P	25ng/rxn	5.98	2.13E-03	3.55E-04	4.92	8.23E-04	18.72
L747_E749	25ng/rxn	94.88	5.45E-02	5.74E-04	7.72	8.14E-05	1.85

Table 3.1: **Enzyme kinetic parameters and erlotinib binding affinity for EGFR WT and ex19del variants. Data produced by SignalChem and analyzed by Patrick Finneran and Benjamin P. Brown.**

Our simulations create structural hypotheses for these differences: First, ex19del variants make distinct hydrogen bonding interactions at the $\beta 3\alpha$C interface (Figure 3.8A – D). E746_A750 places S752 at the $\beta 3\alpha$C i+2 position (Figure 3.1D) such that the sidechain donates a H-bond to the F723 backbone and is simultaneously stabilized as a H-bond acceptor from the K754 backbone (Figure 3.8B). Neither E746_S752>V nor L747_A750>P, both of which place a proline at i+2, can make this H-bond (Figure 3.8C, D). Quantitation of apo-state H-bonding supports this observation, suggesting the glycine-rich loop is more tightly coupled to the $\beta 3\alpha$C-loop in E746_A750 (Figure 3.8E). These data, together with previous crystallographic (Brown et al., 2017) and kinetic (Yosaatmadja et al., 2015) studies of EGFR L858R, suggest generally that tight coupling of the $\beta 3\alpha$C-loop to the glycine-rich loop in αC-helix-stabilizing oncogenic mutants leads to reduced ATP binding affinity.

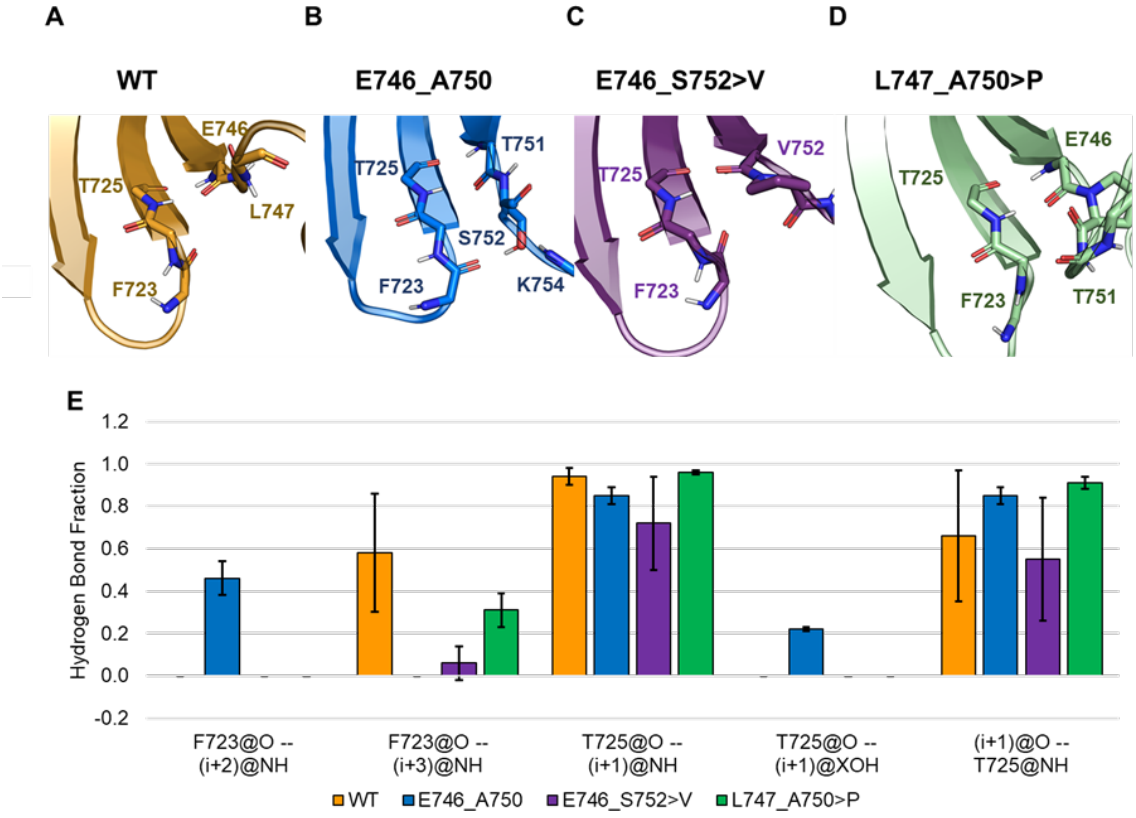


Figure 3.8: Conventional MD simulations demonstrate ex19del $\beta 3\alpha$C hydrogen bond networks. Apo-state conventional MD simulation snapshots of $\beta 3\alpha$C hydrogen bond networks in (A) WT, (B) E746_A750, (C) E746_S752>V, and (D) L747_A750>P. (E) Quantification of hydrogen bond stability of select $\beta 3\alpha$C hydrogen bonds at the interface. Hydrogen bonds are defined by donor/acceptor heavy atom distances of 3.5 and angles between 135 and 180 degrees. Quantifications are based on three independent trials of 4.0 us apo-state simulations of each system starting from the active state.

3.2.7 New therapeutic strategies may be required to maximally inhibit E746_S752>V-mediated disease

We previously identified the TKI neratinib as a potential therapeutic agent for certain forms of HER2/HER3-mutant cancers in which pan-TKI resistance seems to be associated with enhanced ATP binding affinity (Hanker et al., 2021). Employing the same strategy for neratinib as we did for osimertinib, we performed MD simulations and subsequent MMPBSA binding free energy estimates of ex19dels complexed with neratinib. Our simulations suggest that all of the tested ex19dels reversibly bind neratinib better than osimertinib, but that E746_S752>V has a better neratinib binding energy than E746_A750 or L747_A750>P (Figure 3.9A). Evaluation of neratinib function inhibition in Ba/F3 cells stably transfected with E746_A750, E746_S752>V, or L747_A750>P demonstrate a complete ablation of pEGFR in E746_S752>V and L747_A750>P at 30 nM. Phosphorylation is largely reduced in E746_A750 at 30 nM and completely ablated at 150 nM (clinical-relevant dose, Figure 3.9B, C). We also observed that neratinib effectively reduced pEGFR in lung adenocarcinoma cell lines expressing E746_A750, E746_S752>V, or L747_A750>P (Figure 3.9D – F).

3.3 Discussion

Considerable effort has been paid over the last decade to define the molecular mechanisms of oncogenesis and acquired drug resistance in the most commonly occurring EGFR mutations, specifically L858R and "exon 19 deletion" (Carey et al., 2006; Mulloy et al., 2007; Zhang et al., 2006; Shan et al., 2012; Yun et al., 2008b; Red Brewer et al., 2013). These efforts resulted in development of more effective targeted therapies, including today's first-line therapy for EGFR-mutant NSCLC, osimertinib (Yver, 2016). Despite next-generation sequencing has identified the heterogeneity in the various distinct ex19del variants, the allele-specific mechanisms have not been extensively evaluated. The potential reduced likelihood of non-canonical ex19del variants developing T790M or C797S in response to first or third generation TKI, respectively (Zhao et al., 2020; Zheng et al., 2020), may be because a number of these variants have reduced TKI sensitivity in the setting of higher ATP binding affinity. Indeed, both our group (Brown et al., 2019a) and others (Fassunke et al., 2018) found the G724S resistance mutation to occur preferentially to C797S in E746_S752>V and related non-canonical variants in response to osimertinib. However, at present, there has not been a systematic evaluation of patient responses to different TKI based on the specific ex19del variant present in tumor. Thus, it is imperative that we investigate individual ex19del variants pre-clinically to ultimately help guide clinicians in therapeutic decision-making.

Here, we have performed detailed computational, biophysical, and biochemical analyses on a diverse subset of some of the most frequently occurring ex19del variants: E746_A750, E746_S752>V, and L747_A750>P. Our data show clear differences in the activation profiles and TKI sensitivities of these ex19del variants with

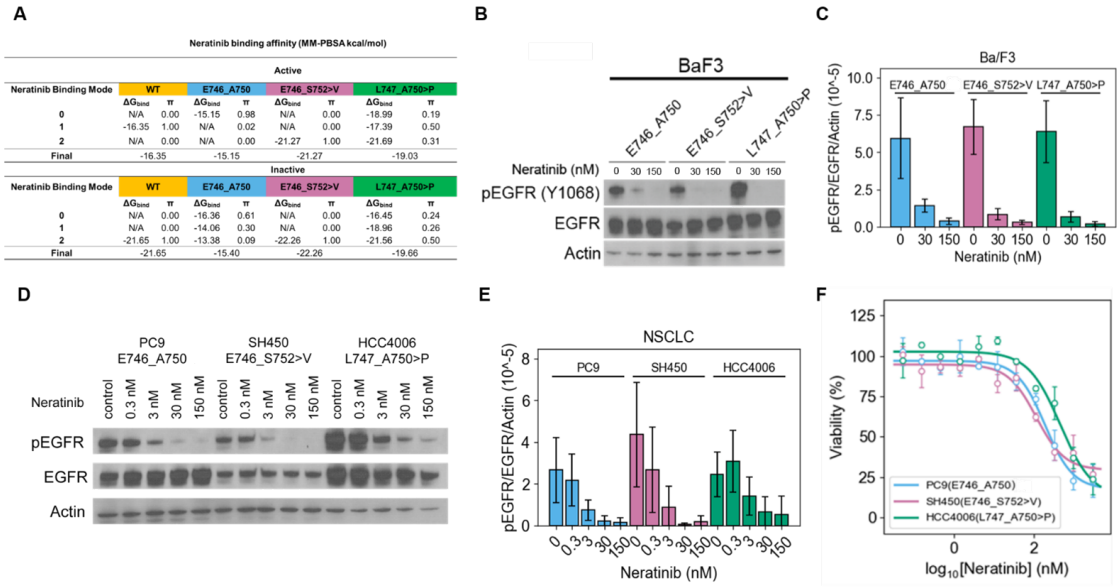


Figure 3.9: Neratinib effectively inhibits E746_S752>V. (A) Neratinib binding affinities for each ex19del variant and WT from simulations starting in the active and inactive states. Three binding modes of neratinib distinguished by the dihedral conformations of the hydroxymethyl pyridine were distinguished with a simple Markov state model. MM-PBSA was not performed if the stationary distribution for a state was estimated at less than 0.05 or the model failed to pass a Chapman-Kalmogorov test for three or two states. Binding energies are computed as the average MM-PBSA energies of 1000 randomly selected frames from the corresponding MSM cluster. For each EGFR variant, six simulations of 2.0 us each were performed such that there were three each from the active and inactive states. (B) Ba/F3 cells were stably transfected with different EGFR ex19del variants and treated with increasing concentrations (0, 30, or 150 nM) of neratinib. Cellular lysates were probed with the indicated antibodies to measure phosphorylation. (C) Quantification of Ba/F3 neratinib inhibition Western blots are represented as the average grayscale ratio of pEGFR/EGFR/Action +/- standard deviation across three independent biological replicates. (D) Ba/F3 cell Lung adenocarcinoma cell lines expressing E746_A750 (PC9), E746_S752>V (SH450), or L747_A750>P (HCC4006) were treated with increasing concentrations (0, 0.3, 3, 30, or 150 nM) of neratinib. Cellular lysates were probed with the indicated antibodies to measure phosphorylation. (E) Quantification of lung adenocarcinoma cell line neratinib inhibition Western blots are represented as the average grayscale ratio of pEGFR/EGFR/Actin+/- standard deviation across three independent biological replicates. (F) Cell viability assays performed in lung adenocarcinoma cell lines stably expressing E746_A750 (PC9), E746_S752>V (SH450), or L747_A750>P (HCC4006) with neratinib. Data and illustrations for figure panels B - F produced by Yun-Kai Zhang, Yingjun Yan, Zhenfang Du, Jiyoon Kim, and Christine M. Lovly.

potential structural correlates. Specifically, our data suggest that the ligand dependency of receptor activation differs between ex19dels. The L747_A750>P mutant displayed robust αC-helix stabilization from a proline-locked tight turn in MD simulations that translated to ligand-independent dimerization and increased in vitro activity in experiments. We also observed that E746_S752>V and L747_A750>P were less sensitive to inhibition by TKI than E746_A750, with E746_S752>V displaying the least sensitivity. We were unable to attribute this effect to binding affinity based on MD simulations of osimertinib or ADP-Glo inhibition assays for erlotinib. Instead, our data suggest a role for variable ATP binding affinity as a potential mediator of these differences in TKI sensitivity. It has previously been observed that some oncogenic EGFR mutations can modulate ATP binding and TKI sensitivity (Carey et al., 2006; Mulloy et al., 2007; Yun et al., 2008b; Yoshikawa et al., 2013).

Collectively, our data demonstrate that ex19dels are a heterogenous group of oncogenic variants. EGFR WT is a monomer in the absence of ligand and stimulated by extracellular EGF to form dimers and multimers/oligomers (Figure 3.10, yellow). The most frequently occurring ex19del oncogenic mutants, such as E746_A750, increase the propensity for dimerization by stabilizing the acceptor KD (Figure 3.10, blue). These "classical super acceptors" (Zhang et al., 2006; Red Brewer et al., 2013) are ligand-dependent and have lower ATP binding affinity (Carey et al., 2006), increasing their sensitivity to TKIs with lower reversible binding affinity, such as osimertinib (Schwartz et al., 2014). Our simulations and TKI sensitivity data suggest that a subset of ex19del variants, such as E746_S752>V and L747_A750>P, are "tight ATP binders" (Figure 3.10, pink). These are characterized by ATP binding affinities higher than that of classical super acceptors, making them more resistant to ATP-competitive TKIs, reminiscent of T790M-comutant EGFR. Unlike e.g., L858R/T790M, the apparent inhibitor potency does not differ from the single oncogenic variant e.g., L858R by several orders of magnitude (Yun et al., 2008b); instead, the difference is 20x. Thus, we distinguish differences in sensitivity from differences in resistance. Finally, another subset of ex19dels, such as L747_A750>P, are characterized by enhanced dimerization propensities greater than that of supper acceptors. These "hyper acceptors" display increased functional activation and exist as ligand-independent dimers (Figure 3.10, green). The ligand-independent activity of hyper acceptors suggest that some oncogenic variants may be activated via "inside-out" dimerization.

Based on our proposed model, L747_A750>P is both a hyper acceptor and a tight ATP binder, while E746_S752>V is a classical super acceptor and potentially a tight ATP binder. E746_A750 is strictly a classical super acceptor. We hypothesize that ex19del variants exist along a spectrum of dimerization propensities and ATP affinities. Based on predicted structural similarity to the mutants studied in depth here, we propose initial classifications of the rarer ex19del variants identified in AACR GENIE along this spectrum. We anticipate that additional functional characterization of ex19del variants along these axes will allow more

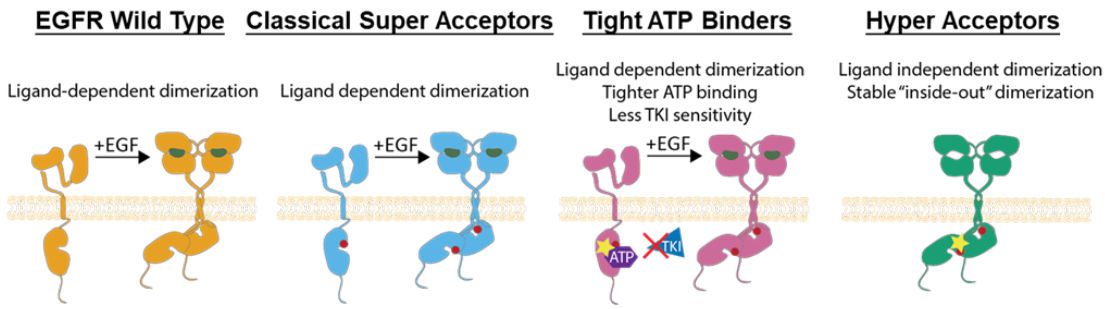


Figure 3.10: Model of ex19del allele-specific functional differences and strategy for inhibition. Discretized classification scheme for EGFR ex19del variants: non-oncogenic with ligand-dependent activation (orange; WT); oncogenic super acceptor with ligand-dependent activation (blue; E746_A750, E746_S752>V); tight ATP binder (pink; E746_S752>V, L747_A750>P); oncogenic hyper acceptor with ligand-independent activation (green; L747_A750>P).

personalized treatment of ex19del NSCLC patients.

Generally, our data lead us to suggest that treatment of ex19del variants may require unique consideration of the variant's functional properties. For example, we speculate that mutations with enhanced ligand-independent dimerization would be less amenable to EGF-blocking antibody / TKI combination therapies than classical super acceptor-like variants. We also suggest that for ex19dels with high ATP binding affinities, the use of covalent TKIs with higher reversible binding affinities may be necessary to overcome reduced TKI sensitivity, such as neratinib or mobocertinib. Alternatively, because increasing the reversible binding affinity on covalent inhibitors can reduce mutant selectivity and cause undesirable side-effects, recognition of tight ATP binding ex19dels may motivate the design of mutant-selective PROTACs or allosteric inhibitors.

On the basis of predicted structural similarity to these three ex19del variants and existing structures of EGFR WT and L858R, we hypothesize functional classifications of the remaining variants from AACR GENIE. Aside from the rarity of most of the ex19del variants we identified in AACR GENIE, only 50% of patients even receive standard-of-care biomarker testing for targetable variants in EGFR and other genes (Robert et al., 2021). Biochemical characterization and stratification into actionable groups is therefore of considerable interest for providing the best possible clinical care to patients with these mutations. To facilitate future comparisons and refinement of our proposed framework, we have made our computational structural models of these variants publicly available on GitHub.

This study is not a comprehensive guide to EGFR ex19del variants. We hope that subsequent work expands upon this study to better characterize uncommon ex19dels. While in silico modeling can provide useful insight to generate hypotheses, it can be limited by factors such as the quality of the predicted structures, the short simulation timescales currently accessible, the start- and end-state dependency of umbrella sampling

simulations, and the simplification of the system from transmembrane dimers/multimers to monomeric intracellular KDs. Similarly, in vitro data in the absence of structural characterization and dynamical insight can make it challenging to generalize findings and perform rational drug design. We anticipate that continued characterization of ex19del structures through experimental structural biology, detailed kinetics studies, and receptor signaling/crosstalk studies will be an important next step in ongoing efforts to design new treatment strategies for patients with EGFR-mutant NSCLC.

3.4 Materials and Methods

3.4.1 Tyrosine kinase inhibitor source and preparation

Inhibitors were purchased from Selleck Chemicals.

3.4.2 Cell culture

Ba/F3 cells (DSMZ), PC9 (ATCC), SH450 (ATCC), and HCC4006 (ATCC) were cultured in RPMI 1640 with L-glutamine (Mediatech) supplemented with 10% heat-inactivated FBS (Thermo Fisher Scientific), penicillin (100 U/mL; Thermo Fisher Scientific), streptomycin (100 μg/mL; Thermo Fisher Scientific), and IL3 (1 ng/mL; Thermo Fisher Scientific) until retroviral transduction and subsequent IL3 withdrawal. Cells were grown in a humidified incubator with 5% CO2 supply at 37°C. Mycoplasma contamination was evaluated routinely during cell culture using a VenorGeM Mycoplasma Detection Kit (Sigma-Aldrich).

3.4.3 Generation of EGFR-expression constructs and generation of Ba/F3 cell lines

pBabe plasmids with EGFR ex19del mutation encoding cDNAs (EGFR E746_A750, EGFR E746_S752>V, EGFR L747_A750>P) and EGFR WT were purchased from Addgene. The empty pBABE-puro retroviral vector or pBABE-EGFR mutants were transfected, along with the envelope plasmid pCMV-VSV-G (Cell Biolabs, San Diego, CA, USA), into cells Plat-GP packaging cells (Cell Biolabs). 48 hours after transfection, viral media was collected, and the debris were removed by centrifugation. For each separate transduction, 1 x 106 Ba/F3 were re-suspended in the viral media and supplemented with 10 μg/mL polybrene (Santa Cruz Biotechnology, Dallas, TX, USA). Transduced cells were selected using 2 μg/mL puromycin (Invitrogen). EGFR construct expressions were checked before experiments, and only stable polyclonal populations were used.

3.4.4 Quantitative assessment of cell proliferation during IL-3 withdrawal

Ba/F3 cells that had been transduced with EGFR-expressing constructs, selected with 2 μg/mL puromycin, and growing in media containing 1 ng/mL IL-3 were washed twice with warm PBS to remove IL-3. Cells were re-suspended in media without IL-3 and seeded in 96-well imaging plates at a density of 3,000 cells/well.

Cells were periodically scanned in IncuCyte® ZOOM every 6 hours using Incucyte® Nuclight Rapid Red Dye for nuclear labeling. Cell doubling values were calculated using the cell counts at each time point divided by the cell counts at start time point.

3.4.5 Immunoblot and antibodies

Antibody EGFR (#2232), pEGFR Y1068, pEGFR Y992, pEGFR Y1184, horseradish peroxidase (HRP)-conjugated anti-rabbit (#7074) were all purchased from Cell Signaling Technology, and the actin antibody (A2066) was purchased from Sigma-Aldrich. For immunoblotting, cells were harvested before or after ligand or drug treatment, washed using PBS, and lysed with RIPA buffer [50 mmol/L Tris HCl (pH 8.0), 150 mmol/L sodium chloride, 5 mmol/L magnesium chloride, 1% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS, 40 mmol/L sodium fluoride, 1 mmol/L sodium orthovanadate, and complete protease inhibitiors (Roche Diagnostics)]. For signal detection, Western Lightning ECL reagent (PerkinElmer) was used. Phosphorylated bands were quantified using ImageJ.

3.4.6 Viability assays

Experiments were conducted in the Vanderbilt High Throughput Screening Facility. Cells were seeded at approximately 800 cells per well in 384-well plates using Multidrop™ Combi Reagent Dispenser (Thermo Scientific). Medium containing different drug concentrations were prepared using a column-wise serial 3X dilution in 384-well plates using a Bravo Liquid Handling System (Agilent) and were added to the cells. Cell viabilities are obtained using CellTiter-Blue® Cell Viability Assay (Promega).

3.4.7 Statistical analysis

All experiments were performed at least three time and the difference were determined by ordinary one-way ANOVA using GraphPad Prism 9.2.0. Difference was considered significant when $p < 0.05$.

3.4.8 Enzymatic analysis

EGFR WT (#E10-112G, lot J3837-8), E746_A750 (#E10-122JG, lot O3886-10), L747_A750>P (#E10-12MG, lot G1200-3), and L747_E749 (#E10-12LG, lot G1344-5) were purchased from SignalChem. The Promega ADP-Glo™ kinase assay kit was used to quantify the amount of ADP produced by each EGFR variant in 1XBFA buffer and in the presence or absence of erlotinib at varying concentrations. Poly(4:1 Glu, Tyr) at a concentration of 0.2 μM was used as the peptide substrate. Reactions were performed at room temperature for 40 minutes each at varying ATP concentrations: 3.125, 6.25, 12.5, 25, 100, 500 μM. Reactions were performed on 384-well plates with each ATP concentration performed in duplicate. Following incubation for 40 minutes, the Promega ADP-Glo™ reagent is utilized to quench the enzymatic reaction and

remove residual ATP. The kinase detection agent provided with the assay kit is subsequently used to convert product ADP back into ATP and measure luminescence from the ATP-powered luciferase/luciferin reaction. ATP Km and erlotinib Ki were fit according a mixed model of inhibition using GraphPad Prism 9.3.1.

3.4.9 Pulsed Interleaved Excitation Fluorescence Cross-Correlation Spectroscopy (PIE-FCCS)

FCCS data were taken on a customized microscope system to introduce pulsed interleaved excitation (PIE) and time-correlated single photon detection as shown in previous works (Huang et al., 2016). A supercontinuum pulsed white laser (9.74 MHz repetition rate, SuperK EXW-12 NKT Photonics, Birkerød, Denmark) was split into 488 nm and 561 nm using filters and mirrors for the excitation of eGFP and mCherry, respectively. The 50 ns time delay for PIE was introduced by directing the splitted beams through two different-length optical fibers (Kaliszewski et al., 2018; Comar et al., 2014). The beams were cleaned, overlapped, and directed to the microscope. A 100X TIRF oil objective (Nikon, Tokyo, Japan) was used for the excitation beam focus and fluorescence emission collection. NIST traceable fluorescein (50 nM; Thermo Fisher Scientific) was used for optical path alignment, and a short, fluorescent-tagged DNA was used as both alignment and as fc value control. Previously published negative and/or positive controls (Kaliszewski et al., 2018; Comar et al., 2014) were tested before the experiment for data quality control and comparisons of the fit parameters. The overlapped excitation beams were focused on to the fluorescently tagged EGFR (WT or mutant)-transfected COS7 cell membrane. The z axis scan was done to ensure that the laser beam was focused on the flat, peripheral membrane area. One 60-second data acquisition was taken per area per cell. The emitted fluorescence was collimated, separated, and filtered before focused onto single-photon avalanche diodes (Micro Photon Devices, Bolzano, Italy) independently. A time-correlated single photon counting module (Picoharp 300, PicoQuant, Berlin, Germany) recorded the time-tagged photon counts for each channel. For analysis, the time-tagged photon counts were divided into six 10-second acquisitions, binned, and gated for channel differentiation. Auto- and cross-correlation curves corresponding to each species were calculated and generated using a custom MATLAB script. Curves of each acquisition per area were filtered, averaged, then fitted to a single component, 2D diffusion model. The averaged and fitted auto-correlation curves show the average dwell time (τD) that we use to calculate the effective diffusion coefficient, Deff = ωo2/4τD. The amplitude of the curves can be used to calculate the local concentration of the diffusing receptors in the detection area. Using the cross-correlation curve, we can calculate cross-correlation values (fc) that indicate the degree of oligomerization. Based on the fc calibration using live cell control system, expected fc value for a monomer-dimer equilibrium is 0.10 to 0.15. Higher fc values indicates higher order oligomerization (Kaliszewski et al., 2018; Comar et al., 2014).

3.4.10 Computational modeling

Structural modeling of proteins was carried out using the Rosetta v.3.12 package (Song et al., 2013). Molecular dynamics simulations were performed with Amber18 utilizing the Amber ff14SB and GAFF2 forcefields for proteins and ligands, respectively (Case et al., 2018). We estimated protein-ligand binding free energies using the MMPBSA.py package in AmberTools18 (Miller et al., 2012). RMSD, atom-atom distances, and dihedrals angles were obtained using CPPTRAJ in AmberTools (Roe and Cheatham, 2013). Markov modeling analysis was performed with PyEMMA2 (Scherer et al., 2015). The initial structure of osimertinib was taken from PDB ID 4ZAU (Yosaatmadja et al., 2015). The initial structure of neratinib was obtained PDB ID 3W2Q (Sogabe et al., 2012). The structures were geometry optimized using Gaussian 09 revision D.01 at B3LYP/6-31G(d) level of theory and the electrostatic potential of the optimized structures computed with HF/6-31G(d) in the gas phase. Atomic partial charges were fit with the restrained electrostatic potential (RESP) algorithm in AmberTools. ATP parameters were developed previously (Meagher Kristin et al., 2003) and coordinates initialized from PDB ID 2ITX. For protein-ligand complexes of variants with osimertinib, neratinib, or ATP, we utilized the above PDB structures for ligand placement.

3.4.11 EGFR ex19del structural modeling

We first built structural models of the 60 ex19del variants identified in AACR GENIE with RosettaCM using the REF2015 score function (Alford et al., 2017). As templates, we selected the active state EGFR WT structures from PDB IDs 2ITX and 2GS6. We also used the active state model of L858R from PDB ID 4I20. We also included as templates the MD equilibrated structural models of E746_A750 and E746_S752>V we made for our prior study (Brown et al., 2019a). We generated 5,000 RosettaCM models for each variant. The best scoring variant from each was simulated with GaMD for 1.0 us (60.0 μs total). GaMD simulation trajectories were clustered with DBSCAN in CPPTRAJ based on $\beta 3\alpha$C loop RMSD. Each variant was subsequently remodeled with RosettaCM to generate 10,000 more models using the DBSCAN cluster centers as additional templates alongside the prior templates. The best scoring model in round two is the final model. Active state L747P was modeled as a point mutation using the Rosetta PackRotamersMover and FastRelax mover starting from EGFR WT in PDB ID 2ITX. We performed a 1.0 us GaMD simulation on the resulting L747P structure, followed by DBSCAN clustering with CPPTRAJ as above. A representative structure from each cluster was relaxed in Rosetta with progressively ramped-down constraints to the starting coordinates to produce 50 models for each cluster. The best scoring model was carried forward for additional simulations. Inactive state structural models of E746_A750, E746_S752>V, L747_A750>P, L747_T751, L747_P753>S, and L747P were modeled with RosettaCM using the inactive state symmetric dimer EGFR WT in PDB ID 3GT8 as a template.

3.4.12 Conventional MD (cMD) simulations

Each structure was solvated in a rectangular TIP3P box (12 Å buffer) neutralized with monovalent Cl- and Na+ ions (Joung and Cheatham, 2008). Minimization proceeded in three stages: solvent minimization with constraints on solute atoms, solute minimization with constraints on solvent, and subsequently full system minimization without constraints. Each of these stages consisted of 1,000 steps of steepest gradient descent followed by 4,000 steps of conjugate gradient descent. The system was heated in the canonical (NVT) ensemble to 100 K over 100 ps. The system was then heated in the isothermal-isobaric (NPT) ensemble at 1 bar from 100 K to physiologic 310 K over 400 ps. Equilibration was performed in NPT ensemble at 310K for an additional 1000 ps. NPT simulations utilized a Monte Carlo barostat. The temperature was controlled using Langevin dynamics with a collision frequency of 2.0 ps-1. A unique random seed was used for each simulation. SHAKE was implemented to constrain bonds involving hydrogen atoms. Periodic boundary conditions were applied and the particle mesh Ewald (PME) algorithm was adopted for long-range electrostatics with a switching distance of 10 Å. Hydrogen mass repartitioning was employed on solute atoms to allow an integration time step of 4 fs.

3.4.13 Gaussian Accelerated MD (GaMD) simulations

Gaussian accelerated MD (GaMD) is an enhanced sampling method that adds a boost potential to the potential energy surface to accelerate transitions between low-energy states (Miao et al., 2015). The dual boost potential scheme was applied to the system in order to enhance conformational sampling. Systems were equilibrated for 50 ns in cMD. Subsequently, potential statistics for GaMD acceleration were computed from a 10 ns cMD simulation. After addition of the GaMD boost potential, simulations were equilibrated for an additional 50 ns before production. All GaMD simulations were performed in NVT ensemble with a Langevin thermostat and collision frequency of 5.0 ps-1. The upper limit of the boost potential standard deviation was set to 6.0 kcal/mol.

3.4.14 Umbrella sampling and conformational free energy landscapes

Conformational free energy landscapes (FEL) of EGFR WT and ex91del mutants were obtained with constant velocity steered MD (SMD) coupled with Umbrella sampling (US) simulations. The weighted histogram analysis method (WHAM) as implemented by Alan Grossfield (Grossfield) was used to perform final statistical reweighting of the US simulations. SMD simulations of 100 ns were performed with a harmonic bias potential and spring constant of 1000 kcal/mol/Å2. SMD simulations were performed from the active to the inactive state and vice versa using Cα RMSD to the reference coordinates as the collective variable. A minimum of 250 windows were selected from each forward and backward simulation with which to seed

US simulations. Therefore, a total of at least 500 windows per system were used to ensure overlap. A 2D harmonic restraining potential was applied to two CVs for the US simulations. CV1 (y-axis) was defined as the difference in the distance between K860(NZ) – E762(OE1, OE2) and K745(NZ) – E762(OE1, OE2). CV2 (x-axis) was defined as the dihedral angle formed by the Cα atoms of the following residues: D855, F856, G857, and L858. A 2.0 kcal/mol/Å2 spring constant was used for CV1, and a 10.0 kcal/mol/rad2 spring constant was used for CV2. At each umbrella center a 5 ns simulation was performed. The first 1 ns was used for equilibration, and the following 4 ns were used for analysis in WHAM. Lowest free energy pathway (LFEP) analysis completed with the LFEP package freely available from the Moradi Laboratory at the University of Arkansas.

3.4.15 Markov model analysis of molecular dynamics simulations

We constructed hidden Markov state models (MSM) to distinguish between two backbone conformations of the glycine-rich loop at residue positions 723 and 724 for osimertinib binding free energy estimates. We also constructed MSMs to distinguish between up to three dihedral conformations of the hydroxymethyl pyridine ring of neratinib for binding free energy estimates. Each MSM was constructed with 6.0 μs (3 x 2.0 μs for each variant for a given active/inactive state) of MD simulation trajectories where frames were collected every 100 ps. All MSMs were constructed with a lag time of 100 ps. The discretized feature trajectories were clustered using KMeans clustering into 500 microstates. All MSMs were validated with Chapman-Kolmogorov tests. In the case of the neratinib binding mode MSMs, if a receptor-neratinib complex did not sample three binding modes, the MSM was regenerated as a two-state model. If only a single dihedral conformer was effectively sampled throughout all three simulations, it was manually assigned a stationary distribution of 1.0. Otherwise, stationary distributions were estimated from MSMs and used to weight the estimated non-covalent binding free energy.

3.4.16 Binding free energy calculations

The estimated binding free energies between EGFR and TKI (osimertinib or neratinib) was computed with the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) method using the MMPBSA.py program in AmberTools18. From each MSM metastable state, we randomly resampled 1,000 structures to use for binding free energy calculations. For the MM-PBSA calculations, the internal and external dielectric constants were set to 4.0 and 80.0, respectively. The nonpolar component of the solvation free energy was estimated from the solvent accessible surface area with the classical method (INP=1) using default coefficient and offset values. Atomic radii were taken from the parameter-topology file (RADIOPT=0).

CHAPTER 4

Structure-function analysis of oncogenic EGFR Kinase Domain Duplication reveals insights into activation and a potential approach for therapeutic targeting

This chapter is taken from Du, Z.*; Brown, B. P.*; Kim, S.; Ferguson, D.; Pavlick, D. C.; Jayakumaran, G.; Benayed, R.; Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M.; Ali, S. M.; Schrock, A. B.; Zehir, A.; Ladanyi, M.; Smith, A. W.; Meiler, J.; Lovly, C. M. Nature Communications 2021, 12 (1), 138236 (*These authors contributed equally).

4.1 Introduction

Next generation sequencing (NGS) based assays have demonstrated high utility as a diagnostic tool for multiple cancer types (Wheler et al., 2016; Ross et al., 2015; Qin et al., 2019; Disel et al., 2020). Interpretation of tumor genomic test results is often complicated by discovery of 'variants of unknown significance' (VUS), because insufficient evidence is available to confirm whether the variant is a driver (deleterious) mutation (Richards et al., 2015; Li et al., 2017). Previously, we identified a VUS in EGFR that contains a tandem in-frame duplication of exons 18 - 25 in an index patient with metastatic lung adenocarcinoma. Since exons 18-25 encode the entire tyrosine kinase domain, we termed this variant 'EGFR Kinase Domain Duplication' (EGFR-KDD) (Gallant et al., 2015).

The ability to effectively treat patients is rooted in our mechanistic understanding of genomic variants identified via sequencing. The classic example is BRAF mutations, which are detected in numerous tumors (Dankner et al., 2018). There are three classes of BRAF mutations, stratified by mechanism and therapeutic actionability (Dankner et al., 2018; Yao et al., 2015). Generally, class I mutations, most notably V600E, are treated with a B-RAF inhibitor such as vemurafenib or dabrafenib, while class II and III mutations are insensitive to vemurafenib/dabrafenib (Yao et al., 2015). Thus, a primary goal in precision medicine is to identify and mechanistically characterize mutations and translate these findings into clinically actionable therapeutic strategies.

Regarding EGFR, mutations in the kinase domain involving small deletions in exon 19 or point mutation in exon 21 (L858R) have been well described (Pao and Chmielecki, 2010). These mutations increase enzymatic activity by stabilizing the active conformation of the kinase domain to promote receptor dimerization (Shan et al., 2012). Numerous studies have now shown that patients with EGFR kinase domain mutations benefit from treatment with EGFR tyrosine kinase inhibitors (TKIs), whereas patients with tumors containing wild-type EGFR do not derive benefit (Pao and Chmielecki, 2010). Analogously, mutations in the EGFR

extracellular domain (ECD) are detected in patients with glioblastoma but are significantly less sensitive to EGFR TKIs in vitro compared to the EGFR kinase domain mutations found in lung cancer (Vivanco et al., 2012), reinforcing the concept that not all mutations within a given gene can be therapeutically targeted in the same manner. In the case of EGFR-KDD, the entire gene contains wild-type sequence with an intragenic duplication of exons 18-25. The addition of a second kinase domain to the intracellular region of EGFR introduces a potentially significant structural perturbation. The functional and therapeutic implications of this variant remain uncertain. Moreover, the unique biology of this variant may make it a valuable tool in the study of ERBB family members and, more generally, suggests a strategy for the study of kinases.

In the present study, we evaluate the prevalence of KDD in ERBB family members (EGFR/EGFR, ERBB2/HER2, ERBB3/HER3, and ERBB4/HER4) across multiple types of human cancers in order to refine our understanding of KDD as an oncogenic driver. In addition, we combine detailed structural modeling, biochemical assays, and experimental and computational biophysical analyses to understand the mechanism whereby EGFR-KDD aberrantly activates EGFR. Collectively, these complementary approaches suggest that EGFR-KDD is activated through formation of ligand-independent intra-molecular dimers and signaling amplified through ligand-dependent inter-molecular dimers/multimers. Furthermore, we show that inhibition of EGFR-KDD activity is maximally achieved by blocking both intra- and inter-molecular dimerization. These studies have important implications for the treatment of patients whose tumor harbor EGFR-KDD.

4.2 Results

4.2.1 ERBB family KDDs are recurrent in multiple cancer types

To investigate the prevalence of KDD in all ERBB family members, we analyzed clinical NGS data from 237,701 tumor samples within the Foundation Medicine (FMI) database. In total, we identified 799 KDDs in ERBB family members (0.34%, 799/237,701). Of those 799 KDDs, EGFR accounts for 443 (55.4%), ERBB2 217 (27.2%), ERBB3 92 (11.5%), and ERBB4 47 (5.9%). Among the cancers present in the FMI database, ERBB-KDD was found most frequently in glioma (2.4%, 227/9,381 total glioma cases), followed by upper gastrointestinal cancer (upper GI; 0.8%, 89/11,822) and non-small cell lung cancer (NSCLC; 0.2%, 109/48,699). For EGFR-KDD, glioma has the highest frequency (2.4%, 222/9,381), followed by NSCLC (1.4%, 70/48,699) and GI (0.3%, 40/11,822). We observed lower incidences of KDD in ERBB2, ERBB3 and ERBB4 than EGFR, with distributions mirroring those of other observed oncogenic mutations in brain tumors and NSCLC (Brennan et al., 2013; Imielinski et al., 2012; Frattini et al., 2013; Cancer Genome Atlas Research, 2014; Mishra et al., 2017).

We also analyzed 40,165 tumor samples from the Memorial Sloan Kettering Cancer Center (MSKCC) IMPACT database (MSK-IMPACT) (Zehir et al., 2017). These data confirm that KDD occurs most frequently

in EGFR, followed by ERBB2. EGFR-KDD is most prevalent in glioma and NSCLC, while ERBB2-KDD is most prevalent in breast and gynecological cancers (GYN). These distributions are consistent with the observed distributions of other EGFR oncogenic mutations in glioblastoma (Imielinski et al., 2012; Cancer Genome Atlas Research, 2014) and NSCLC (Frattini et al., 2013; Mishra et al., 2017) and other ERBB2 mutations in breast cancer (Nik-Zainal et al., 2016), supporting the notion that specific genes may be genomically altered through a variety of mechanisms in a given tumor context.

The overall frequency of ERBB-KDDs from the two datasets is between 0.58 - 2.4% in glioma, 0.07 - 0.22% in NSCLC, and 0.05 - 0.40% in breast cancer. Differences in detection between the two datasets are likely the result of the different methodologies employed for each dataset to identify KDDs (see Methods). Nevertheless, these data suggest that ERBB-KDD is a recurring oncogenic driver in tumor types known to be dependent on ERBB signaling (lung, breast, etc.).

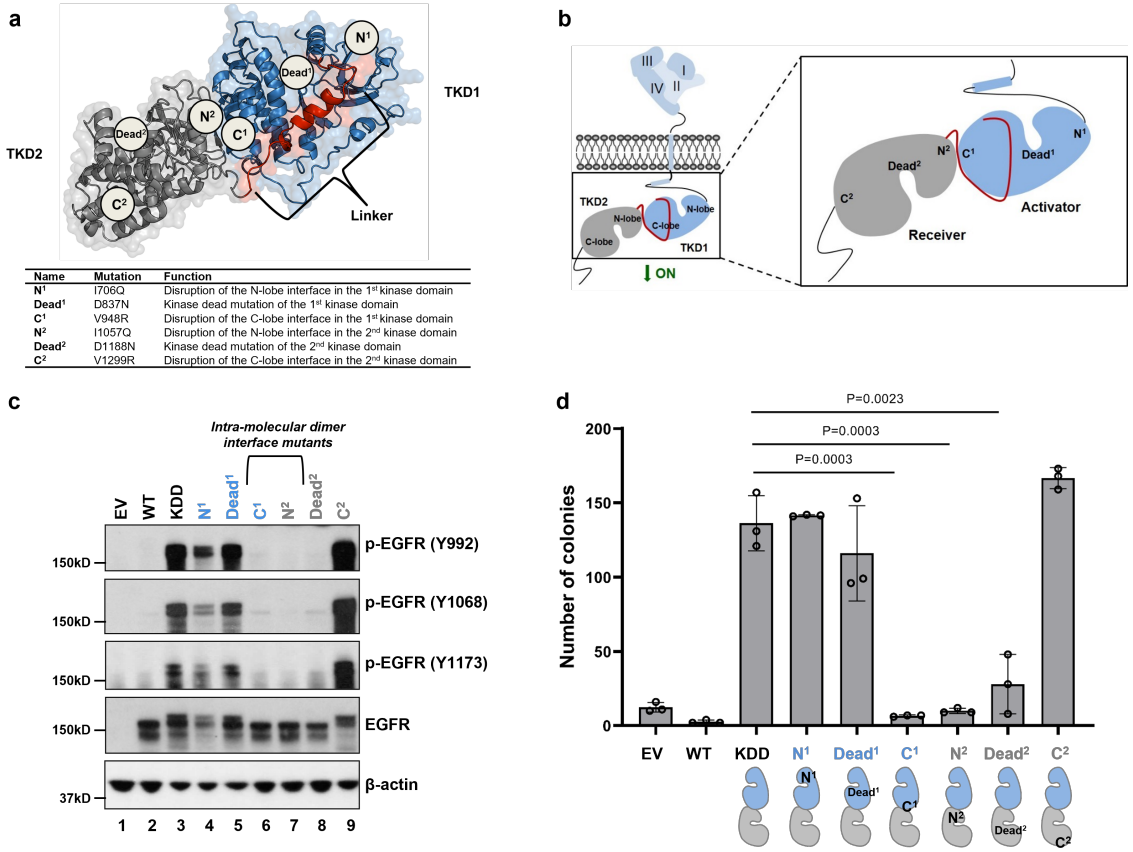
4.2.2 EGFR-KDD is a constitutively active intra-molecular dimer

Even within a single driver gene, the type of mutation that occurs can influence prognosis and drug responsiveness. It is therefore critical to fully characterize the functional consequences of genomic variants in clinically relevant genes. To help us probe the biochemistry of the EGFR-KDD intra-molecular dimer, we leverage core principles of EGFR receptor biology.

ERBB family members are transmembrane tyrosine kinases that possess an extracellular ligand binding domain, a single-pass transmembrane domain, a juxtamembrane (JM) region, an intracellular tyrosine kinase domain (TKD), and a carboxy (C-) terminal tail with multiple tyrosine phosphorylation sites (Lemmon and Schlessinger, 2010). Biochemical and crystallographic studies have shown that activation of EGFR-wild type (WT) involves ligand-induced asymmetric homo- or hetero- dimerization of two TKDs. In the presence of ligand, the C-lobe of one TKD (activator) contacts the N-lobe of another TKD (receiver) to relieve autoinhibition and activate the receiver TKD21. Previous studies of EGFR-WT have identified mutations at the inter-molecular dimer interface that can disrupt dimerization and prevent EGFR-WT enzymatic activity (Zhang et al., 2006).

EGFR-KDD is composed of two intact kinase domains7 (Figure 4.1A). We hypothesized that the forced proximity of the two adjoined kinase domains could form a constitutively active intra-molecular asymmetric dimer in the absence of ligand. To test this hypothesis, we engineered EGFR-KDD constructs with putative intra-molecular dimer disruption mutants (For EGFR mutations, we utilized protein numbering of the human immature EGFR sequence that includes the 24-residue signal sequence) (Figure 4.1A – B): V948R (C1; C-lobe of TKD1) and I706Q (N1; N-lobe of TKD1) in TKD1, and V1299R (C2; C-lobe of TKD2) and I1057Q (N2; N-lobe of TKD2) in TKD2. We also introduced catalytically inactivating mutations (kinase dead) into

each TKD individually (D837N in TKD1 and D1188N in TKD2; Dead[1] and Dead[2], respectively) (Figure 4.1B). We reasoned that these mutants would help us to determine: (1) if EGFR-KDD is catalytically active in the absence of ligand stimulation, (2) the relative orientation of the two intra-molecular kinase domains (i.e. activator vs. receiver), and (3) which of the kinase domains (or both) is catalytically active.



Name	Mutation	Function
N[1]	I706Q	Disruption of the N-lobe interface in the 1st kinase domain
Dead[1]	D837N	Kinase dead mutation of the 1st kinase domain
C[1]	V948R	Disruption of the C-lobe interface in the 1st kinase domain
N[2]	I1057Q	Disruption of the N-lobe interface in the 2nd kinase domain
Dead[2]	D1188N	Kinase dead mutation of the 2nd kinase domain
C[2]	V1299R	Disruption of the C-lobe interface in the 2nd kinase domain

Figure 4.1: Mutations disrupting the potential intra-molecular dimer interface abrogate phosphorylation of EGFR-KDD and anchorage independent growth. a, Ribbon diagram and space-filling model of EGFR-KDD kinase domains. Mutations constructed in this study were labeled. b, Schematic representation of mutations we constructed in this study. We generated point mutations disrupting the potential intra- (C1, N2) and inter-molecular (N1, C2) dimer interface as well as mutations inactivating kinase activity of each kinase domain (Dead[1], Dead[2]). c, YAMC cells stably expressing EGFR-KDD and its mutants. Cells were cultured for 48 hours and then harvested and lysed for analysis. Total EGFR and the auto-phosphorylation at three tyrosine sites were evaluated by western blot. n=3 experiment was repeated independently with similar results. EV, empty vector; WT, EGFR-WT; KDD, EGFR-KDD. d, Soft agar assays were performed in 6 well plates by using YAMC cells. 5,000 cells were seeded in each well and colonies were counted after 4 weeks. n=3 biologically independent samples were examined over 3 independent experiments. Data are presented as mean values ± SD. Statistical differences were analyzed by two-sided unpaired Student's t-test. Data and illustrations for figure panels C and D produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M.

EGFR-KDD and the mutants described above were stably expressed in NR6 (Pruss and Herschman, 1977) (low endogenous EGFR expression) and YAMC (EGFR-/-) (Dise et al., 2008) cells. We evaluated

EGF ligand-independent phosphorylation at EGFR C-terminal tyrosine sites. Ligand-induced dimerization of EGFR-WT results in auto-phosphorylation of its C-terminal tyrosine residues, including Y992 (Walton et al., 1990), Y1068 (Helin et al., 1991) and Y1173 (Helin et al., 1991) (Y1343, Y1419 and Y1524 for EGFR-KDD, respectively). For EGFR phosphorylation sites, we utilized protein numbering of mature EGFR sequence that does not include the 24-residue signal sequence. We observed that EGFR-KDD, but not EGFR-WT, displays phosphorylation of all three tyrosine residues in the absence of EGF ligands (Figure 4.1C, lane 2, 3), indicating that EGFR-KDD is catalytically active without ligand stimulation. We also found that the intra-molecular dimer interface mutants, C1 and N2 (Figure 4.1C, lane 6, 7; Figure 4.2A, lane 6, 7), abolish phosphorylation at all three sites, while N1 and C2 mutants remain phosphorylated in YAMC and NR6 cells (Figure 4.1C, lane 4, 9; Figure 4.2a, lane 4, 9), suggesting that the auto-activation of EGFR-KDD was disrupted by C1 and N2 mutants, rather than N1 and C2 mutants. These data suggest that the N-lobe-mutated TKD1 can activate the C-lobe-mutated TKD2, but not the reverse (Figure 4.1A).

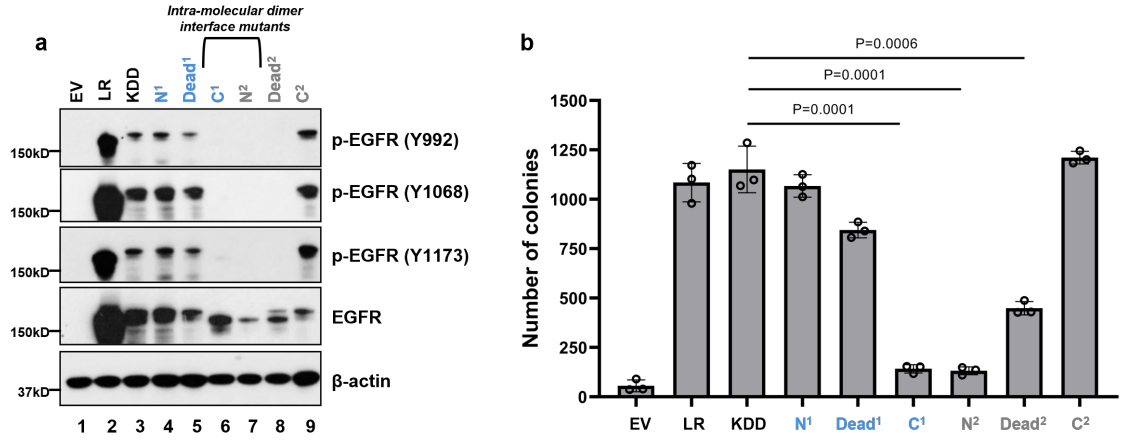


Figure 4.2: Mutations disrupting the potential intra-molecular dimer interface abrogate the auto-phosphorylation of EGFR-KDD activation and anchorage independent growth in soft agar. a, NR6 cells stably expressing EGFR-KDD and its mutants were cultured in serum-free medium for 48 hrs and then cells were harvested and lysed for Western blot. This result is the representative of five independent experiments. b, Anchorage-independent soft agar assays were performed in 6 well plates by seeding 5,000 NR6 in each well. n=3 biologically independent samples were examined over 3 independent experiments. Data are presented as mean values ± SD. Statistical differences were analyzed by two-sided unpaired Student's t-test. EV, empty vector; LR, EGFR L858R mutation. Data and illustrations produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M.

Our catalytically inactive EGFR-KDD TKD2 mutant (Dead^2) failed to autophosphorylate all three tyrosine sites. In contrast, the Dead^1 mutant retained phosphorylation levels comparable to EGFR-KDD in both YAMC and NR6 cells (Figure 4.1C, lane 5, 8 and Figure 4.2A, lane 5, 8). Therefore, in this intra-molecular dimer model, TKD2 functions as the enzymatically active receiver to TKD1, while TKD1 functions as activator to TKD2 (Figure 4.1A).

We further sought to evaluate EGFR-KDD in a phenotypic assay. In both YAMC and NR6 cells, we observed robust colony growth in cells stably expressing EGFR-KDD (Figure 4.1D, Figure 4.2B). We observed that there were comparable numbers of colonies in N1 and C2 mutants compared with EGFR-KDD, while significantly fewer colonies were observed in the intra-molecular dimer-disrupted C1 and N2 mutants (Figure 1D, Figure 4.2B). We also found that Dead^{1}, but not Dead^{2}, could support anchorage-independent growth of YAMC (Figure 4.1D) and NR6 (Figure 4.2b) cells. Therefore, our phenotypic data provide evidence that reduced phosphorylation in the C1 and N2 intra-molecular dimer-disrupted mutants diminish anchorage-independent growth. Taken together, these data are evidence that EGFR-KDD forms a catalytically active asymmetric intra-molecular dimer in the absence of EGF ligand.

4.2.3 Linker contributions to intra-molecular dimer stability

The juxtamembrane B (JMB) domain is an integral component of HER-family homo- and hetero-dimerization. The receiving kinase JMB domain forms specific stabilizing enthalpic contacts in the activator kinase C-lobe (e.g. the hydrophobic residues L688, V689, and L692, and multiple polar contacts) (Red Brewer et al., 2009). Not surprisingly, the JMB residues are highly conserved in HER-family receptors (Figure 4.3A). In EGFR-KDD, the TKD2 JMB is linked directly to the C-terminus of TKD1 (Figure 4.3B). Thus, an important question remained as to whether constitutive EGF-independent activation of EGFR-KDD is the result of (A) sequence-specific structural perturbations to the JMB region, or (B) the sterically imposed forced proximity of TKD1 and TKD2. To address this question, we generated all-atom structural models of EGFR-KDD with Rosetta and molecular dynamics (MD) simulations (Figure 4.4A – C). For comparison, we also modeled the EGFR-WT homodimer.

We measured the per-residue root-mean-square-fluctuations (RMSF) of the linker residues in EGFR-KDD. Our modeling suggests that the linker region corresponding to the JMB is less flexible than the activator C-terminus region, particularly near the N-terminal portion of the JMB (Figure 4.3C – D). Therefore, we hypothesized that the EGFR-KDD JMB forms enthalpically stabilizing contacts at the intra-molecular dimer interface.

To test this hypothesis, we replaced pieces of the linker with unstructured glycine-glycine-serine (GGS) repeats. We substituted (GGS)3 for the JMB part of the linker (KDD-(GGS)3) and (GGS)6 for the activator C-terminus part of the linker (KDD-(GGS)6) (Figure 4.3B). Substitution with (GGS)¬x exchanges sequence-specific contacts with a non-interacting, flexible sequence of matching length28. We transiently transfected the mutants into HEK293 cells and measured EGF-independent receptor phosphorylation via Western blot analysis. KDD-(GGS)3 displays decreased phosphorylation relative to EGFR-KDD, while KDD-(GGS)6 retained similar levels of phosphorylation as EGFR-KDD (Figure 4.3E, lane 3 – 5). Importantly, KDD-

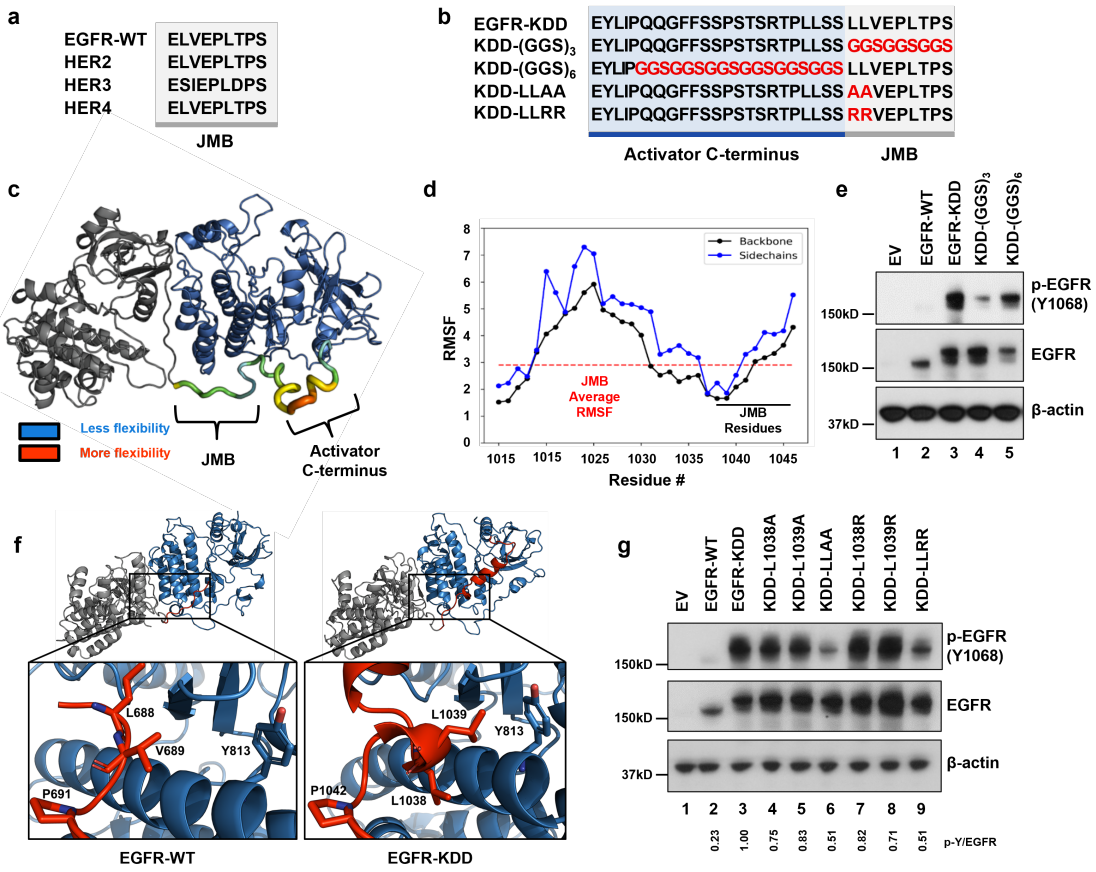
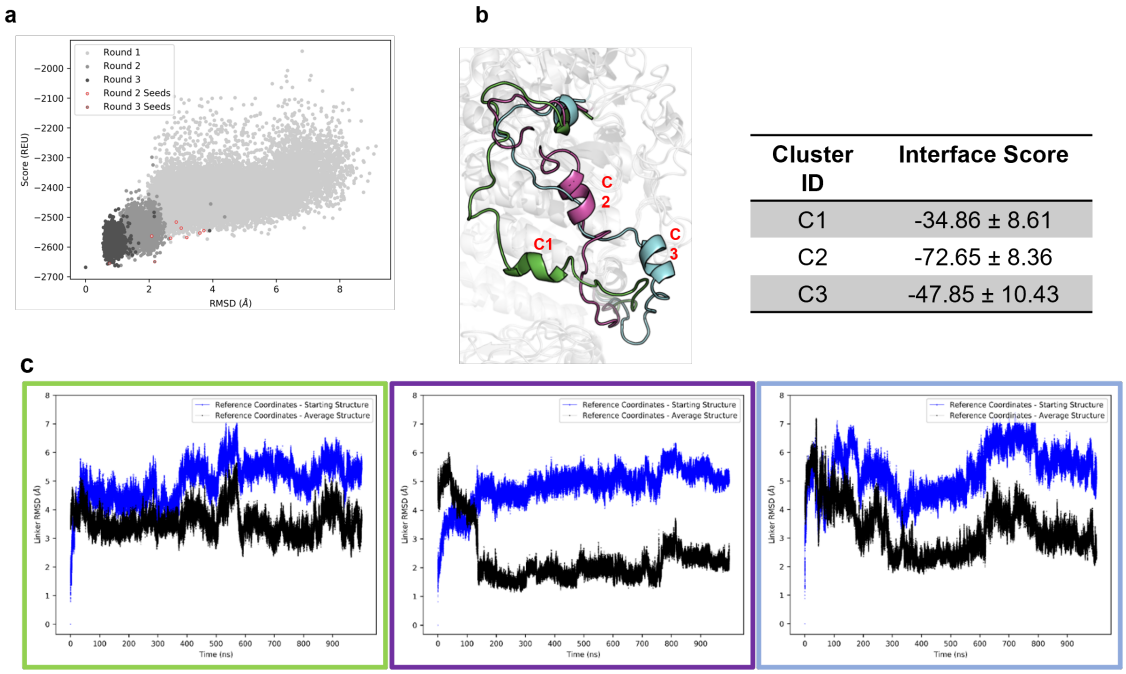


Figure 4.3: The EGFR-KDD linker has distinct enthalpic and entropic contributions to intra-molecular dimer formation. a, Amino acid sequence alignment of EGFR-WT, HER2, HER3, and HER4 JMB domain. b, Amino acid sequence alignment of EGFR-KDD mutants to evaluate linker contributions. Residues in the activator C-terminus kinase domain (TKD1) highlighted in blue (white font). Residues in the receiver JMB domain highlighted in gray (black font). Mutations indicated by red font. c, Per-residue root-mean-square-fluctuation (RMSF) of the EGFR-KDD linker region following an additional 1 μs of MD simulation (post-Rosetta modeling and initial 1 μs MD simulation). RMSF values are mapped onto the structure to indicate regional flexibility. Color gradient and cartoon structure width indicate flexibility. Less flexible = smaller width, colored blue; more flexible = larger with, colored red. d, Graphical representation of per-residue RMSF displays linker residue on x-axis and RMSF on y-axis; black horizontal line indicates JMB residues, red dashed horizontal line indicates average RMSF of JMB residues. e, HEK293 cells transiently transfected with EGFR-KDD or (GGS)n mutants. After 48 hours transfection, cells were collected for western blot analysis. EV, empty vector. f, Detailed structural models of the EGFR-WT homodimer with the JMB domain, and the EGFR-KDD intra-molecular dimer, were generated with Rosetta and refined with 1 μs MD simulations. g, HEK293 cells transiently transfected with EGFR-KDD and different JMB interface mutants. After 48 hours transfection, cells were collected for western blot analysis. p-Y/EGFR, the ratio of phosphotyrosine content at Y1068 to total EGFR expression for each construct relative to EGFR-KDD was shown. EV, empty vector. Data and illustrations for figure panels E and G produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M.



Cluster ID	Interface Score
C1	-34.86 ± 8.61
C2	-72.65 ± 8.36
C3	-47.85 ± 10.43

Figure 4.4: EGFR-KDD intra-molecular dimer model building and refinement. a, Models of the EGFR-KDD intra-molecular dimer were generated with Rosetta. Models from rounds 2 and 3 of the model building process were clustered based on the structure of the linker domain. b, The best scoring model from each of the top three clusters (C1, green; C2, purple; C3, blue) were selected for refinement in Amber18 (left panel). Binding scores for each of the linker conformations (left panel) were computed with MM-GBSA neglecting the entropic contribution to binding (right panel). Frames for inclusion in the MM-GBSA calculation were selected every 100 ps across the entire 1.0 μs trajectory. MM-GBSA scores are represented as mean ± SD. c, Stability of the linker region over each 1 μs MD trajectory was analyzed by computing the RMSD of linker heavy atoms to the position of the conformation at the beginning of the production run (black trace) and the average coordinates from the whole production run (blue trace) for C1 (left panel), C2 (middle panel), and C3 (right panel).

(GGS)3 retains increased activity compared to EGFR-WT (Figure 4.3E, lane 2, 4). Taken together, these data suggest that residues in the JMB portion of the linker contribute to the stability of the EGFR-KDD intra-molecular dimer.

Interestingly, the most stable EGFR-KDD linker model packs two leucine residues (L1038 and L1039) against helices αE and αI, corresponding structurally to residue V689 in EGFR-WT (Figure 4.3F, Figure 4.4B – C, Figure 4.5A – D). EGFR-WT V689 has previously been shown to be necessary for EGFR-WT dimer-dependent phosphorylation27. In agreement with these data, our equilibrated EGFR-WT homodimer preserves the V689 contact (Figure 4.3F, Figure 4.5B). Because L1038 and L1039 were among the most stable residues in the model and correspond structurally to an EGFR-WT residue known to stabilize dimerization (V689), we hypothesized that mutation of these residues would impair EGFR-KDD EGF-independent intra-molecular dimer activity.

To test this hypothesis, we performed site-directed mutagenesis at residues L1038 and L1039. In support of this hypothesis, simultaneous introduction of L1038A/R and L1039A/R (KDD-LLAA and KDD-LLRR) resulted in a substantial reduction in phosphorylation (Figure 4.3G, lane 6, 9). Critically, however, KDD-(GGS)3, KDD-LLAA, and KDD-LLRR all retain increased phosphorylation relative to EGFR-WT (Figure 4.3E, lane 2,4; Figure 4.3G, lane 2, 6, 9). Individual point mutations L1038A/R and L1039A/R do not appreciably reduce phosphorylation; only the combined mutations reduce phosphorylation. Importantly, the sequential leucine residues in the linker are a unique feature of EGFR-KDD resulting from the domain fusion. Altogether, this suggests that despite sequence-dependent JMB contributions to stability, the forced proximity of TKD1 and TKD2 is sufficient for the formation of EGF-independent active intra-molecular dimers. Nevertheless, the linker sequence can provide additional enthalpic stabilization to increase activation.

4.2.4 Ligand induces inter-molecular multimer activity

EGFR-WT activation is achieved through ligand-induced inter-molecular dimerization21. Recent evidence demonstrates that EGFR-WT also forms tetramers and other small oligomers that increase phosphorylation in an EGF concentration-dependent manner29, 30, 31, 32. We wanted to know if EGFR-KDD activity is similarly augmented by EGF-ligand stimulation.

To differentiate between EGFR-KDD activity caused by EGF-dependent inter-molecular dimerization and EGF-independent intra-molecular dimerization, we utilized cetuximab, an anti-EGFR extracellular domain antibody that blocks EGF-mediated EGFR dimerization33. EGF binding leads to inter-molecular dimerization of EGF receptors. Cetuximab prevents EGF binding by blocking the EGF binding site. We stimulated cells expressing various EGFR-KDD constructs with EGF. We found that phosphorylation of EGFR-KDD is dramatically increased in the presence of EGF stimulation (Figure 4.6A, lane 5, 6; Figure 4.6B, lane 5,

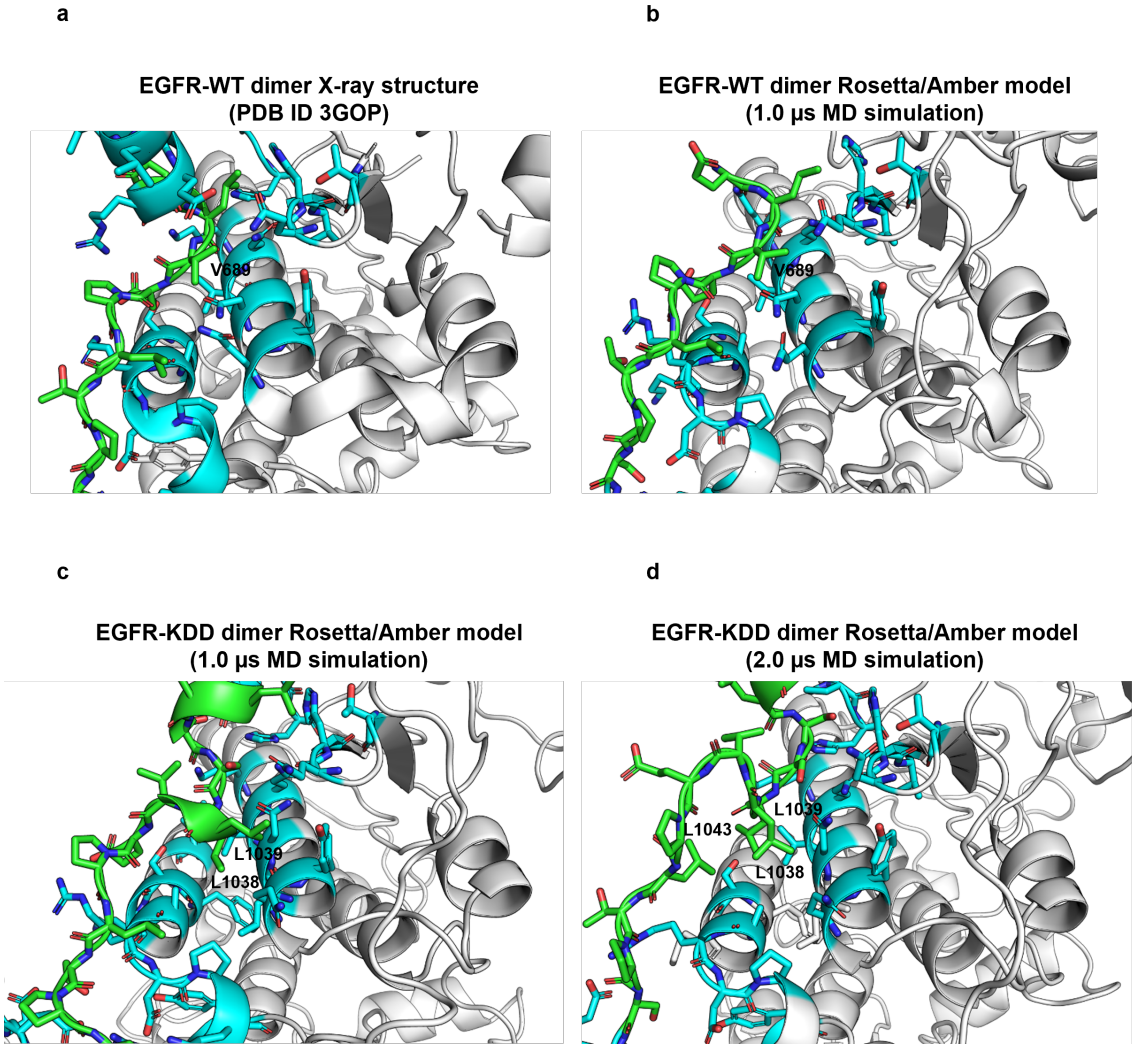


Figure 4.5: Comparison of EGFR-KDD computational models with X-ray structure of EGFR-WT juxtamembrane latch. a, X-ray structure of the EGFR-WT homodimer with juxtamembrane latch; b, Rosetta model of EGFR-WT homodimer with juxtamembrane latch post-equilibration for 1.0 μs MD simulation; c, Rosetta model of EGFR-KDD intra-molecular dimer post-equilibration for 1.0 μs MD simulation; d, Rosetta model of EGFR-KDD intra-molecular dimer post-equilibration for 2.0 μs MD simulation; the receiver kinase domain N-terminal JMB domain is colored green; residues within 6.0 *Å* of JMB are colored blue.

7; Figure 4.7A, lane 5, 6). Addition of cetuximab effectively mitigates EGF-induced phosphorylation of EGFR-KDD (Figure 4.6B, lane 5 - 8, Figure 4.7B, lane 9 - 12). These data suggest that EGF stimulation may promote EGFR-KDD activity through the formation of at least inter-molecular dimers; however, cetuximab does not preclude the formation of dimers entirely.

To further test the hypothesis that EGF stimulation promotes the formation of at least inter-molecular dimers in EGFR-KDD, we administered mAb806 to YAMC EGFR-KDD cells. The mAb806 antibody inhibits EGFR dimerization by binding to extracellular domain II (residues 287–302)34, rather than the EGF ligand binding site in domain III33. Thus, inhibition with mAb806 is highly complementary to similar experiments performed with cetuximab. As expected based on our cetuximab results, we found that mAb806 had no impact on phosphorylation level in the absence of EGF ligand (Figure 4.7c, lane 1, 2, 5, 6) and decreased the level of phosphorylation with EGF-ligand stimulation (Figure 4.7d, lane 3, 4, 7, 8). We also note that phosphorylation was reduced more by cetuximab than mAb806 at approximately equimolar concentrations, consistent with previous reports that the EGFR inhibitory potency of mAb806 is considerably lower than cetuximab35.

We showed above (Figure 4.1C – D) that intra-molecular dimer-disrupted mutants C1 and N2 are not active in the absence of ligand. Unexpectedly, we noticed that EGF-stimulation rescued these mutants, leading to a robust increase in phosphorylation (Figure 4.6A, lanes 11-14; Figure 4.7A, lanes 11-14). We speculated that this could result from either (A) compensatory stabilization of the intra-molecular receiver kinase domains or (B) stabilization of the donor kinase domains during inter-molecular dimerization.

To better understand how inter-molecular dimerization increases EGFR-KDD autophosphorylation, we built template-based structural models of the intracellular portion of the EGFR-KDD inter-molecular dimer based on two proposed EGFR-WT tetramer models: (1) an extension of the inter-molecular dimer model in which each kinase domain is successively asymmetrically docked with another (end-to-end model) (Huang et al., 2016) (Figure 4.6C), and (2) two asymmetric dimers oriented such that the N-lobe and C-lobe of one dimer are in contact with the N-lobe and C-lobe of the other dimer, respectively (side-by-side model) (Needham et al., 2016b) (Figure 4.6D). Other models are possible (e.g. the receiver kinase of one intra-molecular dimer could act as the donor to the receiver kinase of a second intra-molecular dimer). There are currently no experimental structures (e.g. from X-ray crystallography or cryogenic electron microscopy) elucidating the organization of EGFR-WT tetramer or EGFR-KDD inter-molecular dimer. Thus, we built our template-based models of EGFR-KDD intracellular inter-molecular dimer on two published EGFR-WT tetramer models both of which have experimental and computational support.

Our models each consist of two EGFR-KDDs containing an intra-molecular donor (TKD1 or TKD3) and receiver (TKD2 or TKD4) kinase. Both structural models suggest a mechanism for active-state stabilization

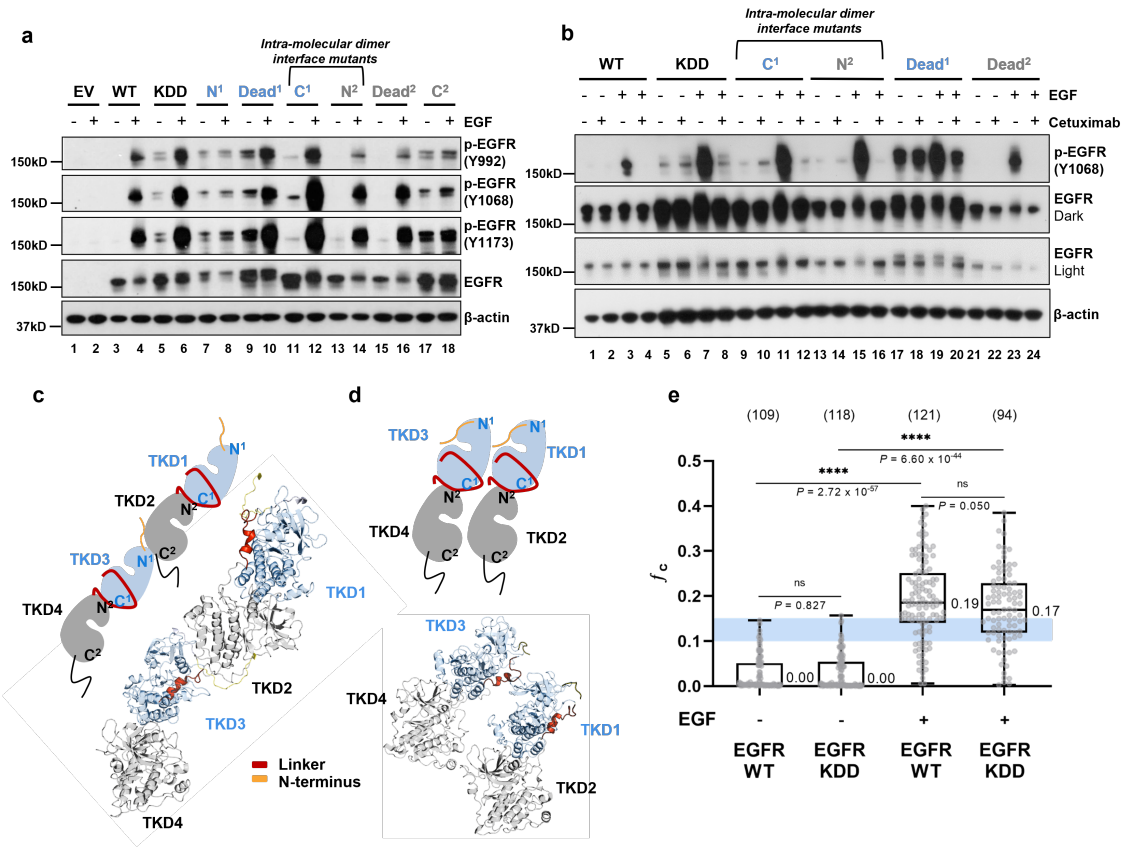


Figure 4.6: EGFR-KDD forms inter-molecular dimers and higher order oligomers after ligand stimulation. a, YAMC cells were cultured in serum-free medium for 12 hours and then treated with 50 ng/mL EGF ligand for 5min. Total EGFR and the autophosphorylation at three tyrosine sites were assessed by western blot. b. YAMC cells were starved for 12 hrs and treated with cetuximab (10 µg/ml in serum-free medium) for 3hrs 45min, and EGF ligand (50 ng/mL in serum-free medium) was added for 15min. The cells were harvested and analyzed by Western blot. WT, EGFR-WT; KDD, EGFR-KDD. c, Template-based structural models of the intracellular portion of the EGFR-KDD inter-molecular dimer based on end-to-end and EGFR-WT tetramer models. d, Template-based structural models of EGFR-KDD inter-molecular dimer based on side-by-side EGFR-WT tetramer model. e, Cross correlation values of EGFR-WT and EGFR-KDD with (+) or without (-) ligand (EGF) stimulation is shown. The blue box indicates the fc value region for dimers. The median values are reported next to the boxplot. Each grey dot represents the averaged acquisition (10 sec, 6 acquisitions) per area per cell. All data points are shown. Numbers in parenthesis above the boxplot are the total number of cells that data were taken on. Data and illustrations for figure panels A and B produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M. Data and illustrations for figure panel E produced by Kim, S. and Smith, A.W.

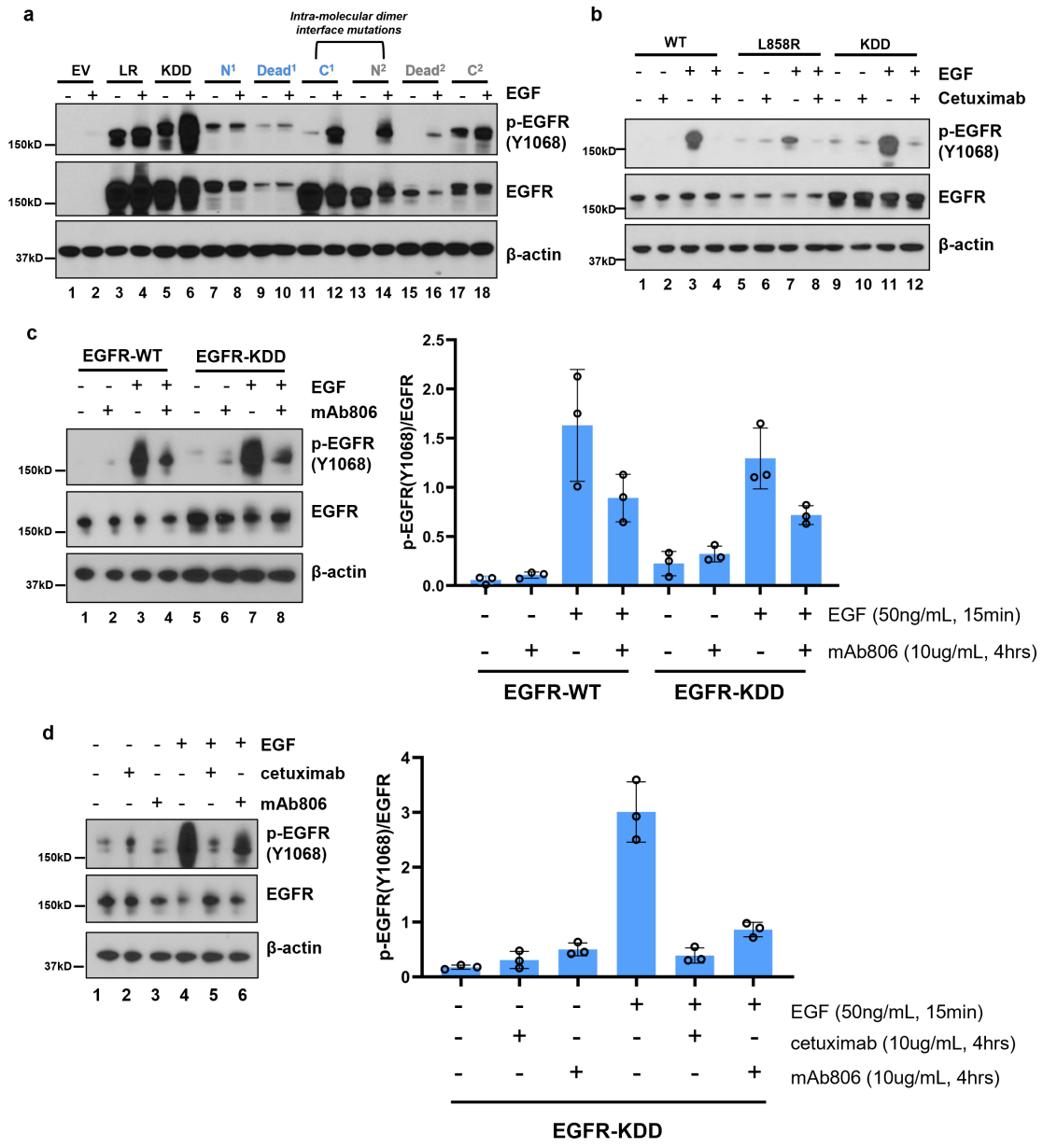


Figure 4.7: Disruption of EGF-induced inter-molecular activation of EGFR-KDD with cetuximab and mAb806. a, NR6 cells were cultured in serum-free medium for 36 hrs and then treated with 50ng/mL EGF ligand for 5min. Total EGFR and the autophosphorylation at three tyrosine sites were assessed by Western blot. b, NR6 cells were starved overnight and treated with cetuximab (10 μg/ml in serum-free medium) for 3hrs 45min, and then were treated with EGF (50 ng/mL in serum-free medium) and cetuximab (10 μg/ml in serum-free medium) for 15min, then cells were harvested for western blot. c, YAMC EGFR-WT and EGFR-KDD cells were starved for 12 hrs and pre-treated with mAb806 antibody (10 μg/ml in serum-free medium) for 3hrs 45min, respectively, and EGF ligand (50 ng/mL in serum-free medium) was added for 15min. The cells were harvested and analyzed by Western blot (left panel). The ratio of phospho-EGFR (Y1068) to total EGFR expression was also shown (right panel). Results represent the mean values of three independent experiments ± SD. d, YAMC EGFR-KDD cells were starved for 12 hrs and pre-treated with cetuximab (10 μg/ml in serum-free medium) and mAb806 antibody (10 μg/ml in serum-free medium) for 3hrs 45min, respectively, and EGF ligand (50 ng/mL in serum-free medium) was added for 15min. The cells were harvested and analyzed by Western blot (left panel). The ratio of phospho-EGFR (Y1068) to total EGFR expression was also shown (right panel). Results represent the mean values of three independent experiments +/- SD. Data and illustrations produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M.

of TKD3 during inter-molecular dimerization (Figure 4.6C – D). In the end-to-end model, active-state stabilization of TKD3 (inter-molecular receiver, intra-molecular donor) could occur by canonical asymmetric dimerization with TKD2 (inter-molecular donor, intra-molecular receiver) (Huang et al., 2016) (Figure 4.6C). In the side-by-side model, active-state stabilization of TKD3 could occur through sterically impaired inactivation by TKD1 (inter-molecular donor, intra-molecular donor) (Figure 4.6D), as observed in the 40 μs MD simulation of the EGFR-WT full-length tetramer model in Needham et al. 2016 (Needham et al., 2016b).

We previously observed that Dead^2 (TKD2 and TKD4 are inactive), but not Dead^1 (TKD1 and TKD3 are inactive), ablates EGFR-KDD activity in the absence of EGF (Figure 4.1C, lane 3, 5, and 8). Here, we see that EGF-ligand stimulation robustly revives phosphorylation in Dead^2 (Figure 4.6A, lane 15, 16; Figure 4.6B, lane 21, 23; Figure 4.7A, lane 15, 16), suggesting active-state stabilization of TKD3 through the formation of at least inter-molecular dimers (Figure 4.6C – D). Less dramatic increases in Dead^1 from baseline intra-molecular dimer phosphorylation are consistent with changes due to ligand-induced EGFR recruitment (Figure 4.6A, lane 9, 10; Figure 4.7A, lane 9, 10). Consistent with these results, pre-administration with cetuximab prevents EGF-dependent phosphorylation of Dead^2 and has only a minor impact on Dead^1 phosphorylation. Taken together, these data suggest that in addition to activation of TKD2 and TKD4 by TKD1 and TKD3, respectively, TKD3 becomes catalytically active in the inter-molecular dimer.

To better characterize the effect of EGF on EGFR-KDD and quantify the extent of EGFR-KDD oligomerization in live cells, we performed two-color pulsed interleaved excitation fluorescence cross-correlation spectroscopy (PIE-FCCS). PIE-FCCS has been previously applied to evaluate EGFR dimerization and multimerization (Huang et al., 2016). For these experiments, the protein of interest was expressed as a mixture of eGFP and mCherry fusions and single, live-cell measurements were recorded and analyzed as described in the Methods section. In the absence of ligand, both EGFR-WT and EGFR-KDD have median cross-correlation (fc) values of 0.00, indicating that they are predominantly monomeric (Figure 4.6E, Figure 4.8B). Stimulation with EGF ligand leads to a significant level of cross-correlation for EGFR-WT (fc = 0.19) and EGFR-KDD (fc = 0.17) (Figure 4.6E, Figure 4.8B), indicating that ligand stimulation induces dimerization and multimerization in both EGFR-WT and EGFR-KDD36, 38. There is no statistically significant difference between EGFR-WT and EGFR-KDD, suggesting that the kinase duplication does not sterically restrict dimerization and multimerization. Taken together, these data demonstrate that EGFR-KDD forms multimers upon ligand binding.

4.2.5 EGFR-KDD directly interacts with ERBB family members

Our biophysical studies demonstrate that EGFR-KDD forms ligand-induced homodimers/multimers. We hypothesized that EGFR-KDD could also heterodimerize with EGFR-WT in the presence of ligand. To

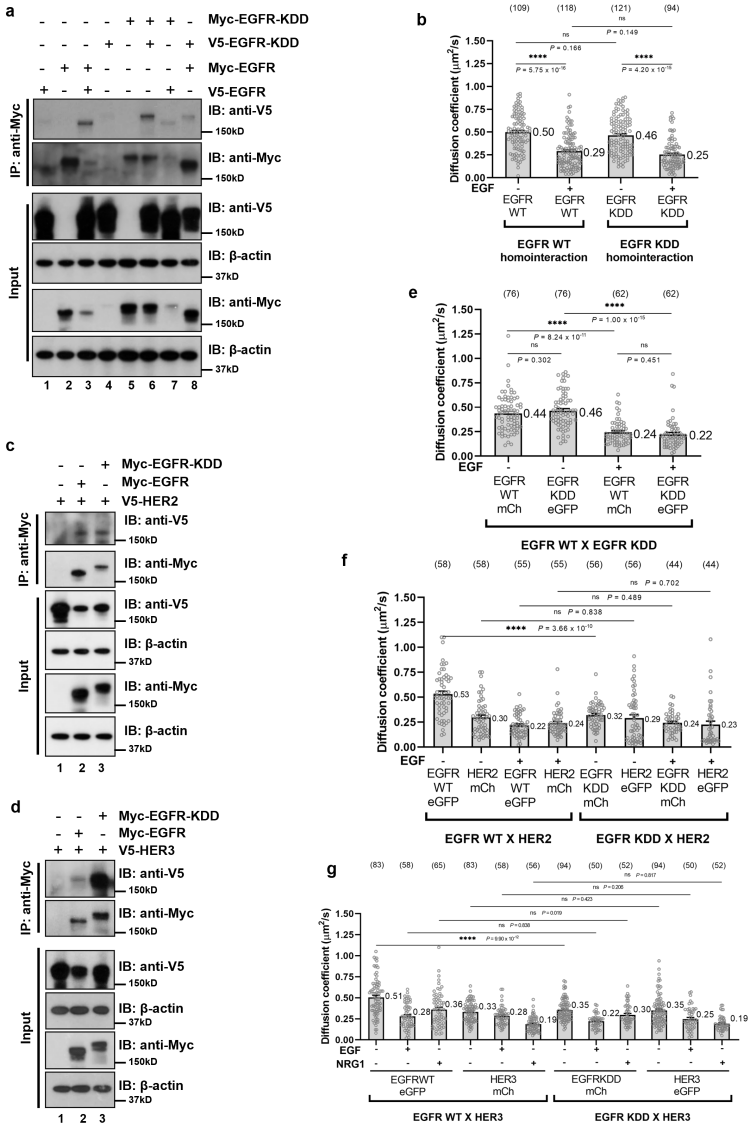


Figure 4.8: EGFR-KDD directly interacts with ErbB family members. a, V5-epitope tagged EGFR-WT and EGFR-KDD was co-transfected with Myc-epitope tagged EGFR-WT and EGFR-KDD in HEK293 cells. Cell lysates were immunoprecipitated by using Myc antibody. Immunoblotting were probed by V5 and Myc antibody. b, Average diffusion coefficient of EGFR WT homodimers with (+) or without (-) ligand (EGF) stimulation is shown. c, V5-epitope tagged HER2 was co-transfected with Myc-epitope tagged EGFR-WT and EGFR-KDD in HEK293 cells. Cell lysates were immunoprecipitated by using Myc antibody. Immunoblotting were probed by V5 and Myc antibody. d, V5-epitope tagged HER3 was co-transfected with Myc-epitope tagged EGFR-WT and EGFR-KDD in HEK293 cells. Cell lysates were immunoprecipitated by using Myc antibody. Immunoblotting were probed by V5 and Myc antibody. e, Average diffusion coefficient of EGFR WT and EGFR KDD mutant with (+) or without (-) ligand (EGF) stimulation is shown. f, Average diffusion coefficient of HER2 and EGFR-KDD mutant with (+) or without (-) ligand (EGF) stimulation is shown. g, Average diffusion coefficient of HER3 and EGFR-KDD mutant with (+) or without (-) ligand (EGF or NRG1) stimulation is shown. Data and illustrations for figure panels A, C, and D produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M. Data and illustrations for figure panels B, E, F, and G produced by Kim, S. and Smith, A.W.

test this hypothesis, we performed co-immunoprecipitation in HEK293 cells with transiently co-transfected Myc-epitope tagged EGFR-KDD/EGFR-WT and V5-epitope tagged EGFR-WT/EGFR-KDD. We observed that V5-epitope tagged EGFR-WT can interact with Myc-epitope tagged EGFR-KDD, and vice versa (Figure 4.9A, Figure 4.8A). We further evaluated potential interactions between EGFR-WT and EGFR-KDD with PIE-FCCS. With the fc values, we can distinguish homo- and heterodimerization, which cannot be assessed with diffusion coefficients alone. EGFR-WT-eGFP and EGFR-KDD-mCherry were simultaneously expressed in COS7 cells. In the absence of EGF ligand, there was no interaction (fc = 0.00). Upon addition of EGF-ligand, there was a significant increase in cross-correlation (fc = 0.22) indicating the formation of heteromeric complexes (Figure 4.9B, Figure 4.8E). The positive cross-correlation is rigorous evidence for heteromeric complex formation, but alone is not sufficient to define the interaction strength or stoichiometry of the complexes. For simplicity we will refer to these complexes as heterodimers as this is the minimal size consistent with positive cross-correlation. In agreement with changes to the fc values, the diffusion coefficients of both EGFR-WT and EGFR-KDD decreased after ligand addition, indicating slower diffusion due to homo- and hetero-dimerization/multimerization (Figure 4.8B and E).

Heterodimerization is especially important for the activation of HER2 and HER3. HER2 has lost the capacity to bind ligands and activates primarily as a receiver kinase domain through heterodimerization with other ERBB family members39, 40. In contrast, the TKD of HER3 has low kinase activity, and HER3 acts as an activator in heterodimers41. We hypothesized that EGFR-KDD can also interact with wild-type HER2 and HER3. To test this hypothesis, we performed co-immunoprecipitation. We transiently co-transfected Myc-epitope tagged EGFR-KDD with V5-epitope tagged HER2-WT and HER3-WT in HEK293 cells. Independent pulldowns with V5 and Myc antibodies demonstrate that EGFR-KDD could interact with HER2 and HER3 (Figure 4.9C, Figure 4.8C – D). Moreover, we observed quantitatively with PIE-FCCS that EGFR-WT and EGFR-KDD heterodimerize with HER2 to a larger extent in the presence of EGF-ligand (fc = 0.10 and fc = 0.16, respectively) than in its absence (fc = 0.00 and fc = 0.06, respectively) (Figure 4.9D, Figure 4.8F). Interestingly, our biophysical data suggest that like EGFR-WT, EGFR-KDD also heterodimerizes with HER3 to a greater extent in the presence of NRG1 than in the presence of EGF (Figure 4.9E, Figure 4.8G). These data demonstrate that EGFR-KDD forms direct interactions with EGFR-WT, HER2 and HER3.

4.2.6 Intra- and inter-molecular dimer activity dual inhibition

The dual nature of EGFR-KDD as an EGF-independent active intra-molecular dimer and as an EGF-dependent active inter-molecular dimer/multimer poses a unique therapeutic challenge. Our computational models and experimental data suggest that the ideal therapy would simultaneously reduce intra- and inter-molecular dimer-mediated activity. One potential treatment strategy is therefore the combination of cetuximab with

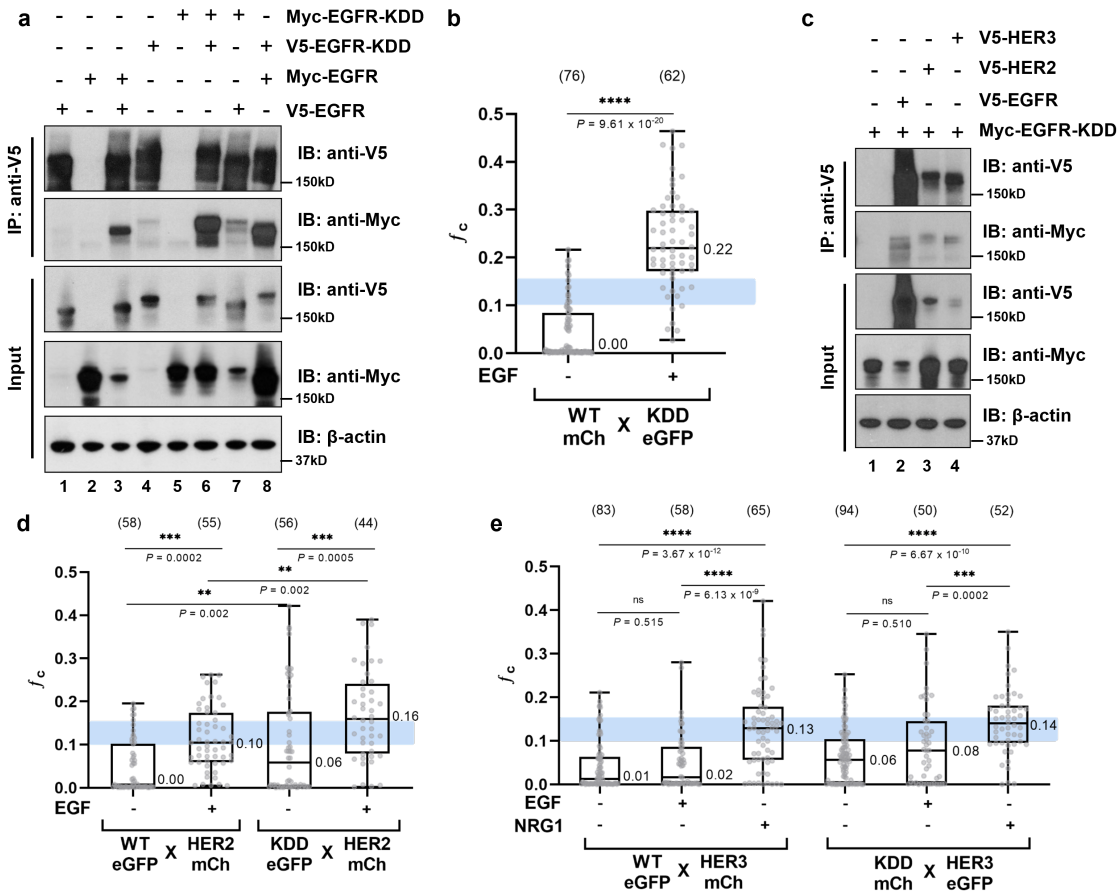


Figure 4.9: EGFR-KDD directly interacts with ERBB family members. a, V5-epitope tagged EGFR-WT and EGFR-KDD was co-transfected with Myc-epitope tagged EGFR-WT and EGFR-KDD in HEK293 cells. After 48 hours transfection, cells were lysed by hypotonic buffer and the cell lysates were immunoprecipitated by using V5 antibody. Immunoblotting were probed by V5 and Myc antibody. b, Cross correlation values of co-transfected EGFR-WT (mCherry-fused) and EGFR-KDD mutant (eGFP-fused) with (+) or without (-) ligand (EGF) stimulation is shown. The light orange box indicates the fc value region for dimers. c, Myc-epitope tagged EGFR-KDD was co-transfected with V5-epitope tagged EGFR-WT, HER2 and HER3 in HEK293 cells. Cell lysates were immunoprecipitated by using V5 antibody. Immunoblotting were probed by V5 and Myc antibody. d, Cross correlation values of co-transfected HER2 (mCherry-fused) and EGFR-KDD mutant (eGFP-fused) with (+) or without (-) ligand (EGF) stimulation is shown. e, Cross correlation values of co-transfected HER3 (mCherry-fused) and EGFR-KDD mutant (eGFP-fused) with (+) or without (-) ligand (EGF) stimulation is shown. For Figure 4.9B, D and E, the median values are reported next to the boxplot. Each grey dot represents the averaged acquisition (10 sec, 6 acquisitions) per area per cell. All data points are shown. Numbers in parenthesis above the boxplot are the total number of cells where data were taken on. Both One-Way ANOVA test and Uncorrected Fisher's LSD test were down to obtain adjusted and individual p values. Source data and statistical analysis are provided in the Source Data file. Data and illustrations for figure panels A and C produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M. Data and illustrations for figure panels B, D, and E produced by Kim, S. and Smith, A.W.

a TKI (here afatinib). Prior pre-clinical literature has suggested that such a combination may be effective in L858R but not Ex19Del42.

The combination of cetuximab with various EGFR TKIs, including gefitinib43 and afatinib44, 45, has been tested in lung cancer patients. In a phase I trial, no responses were observed with the combination of cetuximab plus gefitinib43, and therefore has not been subsequently used in patients. The combination of cetuximab plus afatinib has advanced in the clinic, including a phase I trial (NCT01090011) that included an expansion cohort44, 45. Results from this trial of cetuximab plus afatinib demonstrated that the combination therapy was effective in achieving tumor reduction (as assessed by CT scans using RECIST criteria) in patients with both Ex19Del and L858R EGFR-mutant lung cancer, in contrast to prior pre-clinical data42. Importantly, the combination of cetuximab plus TKI is not FDA-approved because there was no benefit (in terms of PFS, intracranial response, and OS) compared to TKI alone, and thus not standardly used in the treatment of patients with Ex19Del or L858R mutations. The current standard of care for these patients is the mutant-selective EGFR TKI, osimertinib, based on a seminal phase 3 clinical trial46, 47.

In contrast, no pre-clinical study or clinical trial has evaluated antibody/TKI combination vs. either alone in EGFR-KDD patients. Indeed, the index patient for EGFR-KDD described in Gallant et al. 2015 unfortunately only had a partial response to afatinib7. The anti-tumor response was short-lived (7 cycles of afatinib, or approximately 7 months) before the patient developed acquired resistance to afatinib driven by amplification of the EGFR-KDD allele7. Collectively, these observations suggested that more potent EGFR blockade is necessary to overcome the oncogenic activity of EGFR-KDD. Here, we test the hypothesis that combined TKI and cetuximab treatment will reduce EGFR-KDD-mediated phosphorylation in vitro more than either treatment alone.

We treated YAMC cells stably expressing EGFR Ex19Del (E746_A750del), L858R, and EGFR-KDD with afatinib and cetuximab both in the absence and presence of EGF ligand (Figure 4.10A, Figure 4.11a). Importantly, we observed that in both the absence and presence of EGF, afatinib resulted in a near complete ablation of p-EGFR in Ex19Del (Figure 4.10a, lanes 1, 2, 5, 6) and L858R (Figure 4.10A, lanes 9, 10, 13, 14), but substantial residual phosphorylation existed in EGFR-KDD (Figure 4.10A, lanes 17, 18, 21, 22). As expected, cetuximab alone reduced phosphorylation in Ex19Del, L858R and EGFR-KDD in the presence of EGF ligand (Figure 4.10A, lane 7, 15, 23). Notably, the greatest reduction of phosphorylation for EGFR-KDD occurred with the combination of cetuximab + afatinib in the presence of EGF (Figure 4.10A, lanes 21, 22, 23, 24). These data suggest that phosphorylation of EGFR Ex19Del and L858R is abolished by afatinib (TKI) or cetuximab alone, and addition of cetuximab to afatinib does not add substantially more inhibition to the decrease in auto-phosphorylation. Unlike EGFR Ex19Del and L858R, phosphorylation of EGFR-KDD is inhibited by both afatinib and cetuximab as single agent, but the combination treatment yielded more

inhibitory effects.

We also performed viability assays with BaF3 cells stably expressing EGFR-KDD, Ex19Del (E746_A750del) or L858R. First, we evaluated Ba/F3 cell growth in serum starved (0.5% fetal bovine serine; FBS) conditions to minimize EGF activation (Figure 4.11b). At 0.5% FBS, cetuximab maximally exhibited 40% inhibition of EGFR-KDD, 80% inhibition of Ex19Del, and almost 100% inhibition of L858R cell viability (Figure 4.10B). These data are consistent with a model in which EGFR-KDD retains an active intra-molecular dimer in the absence of EGF stimulation (Figure 4.1) and previously published models of Ex19Del and L858R in which intrinsic αC-helix stabilization transforms them into dimer-dependent "super acceptor" kinases. Indeed, progressively higher concentrations of FBS and the addition of exogenous EGF resulted in stable or increased viability of all mutants in the presence of cetuximab, though EGFR-KDD proved to be the least inhibited (Figure 4.10C, Figure 4.11C – E).

In 0.5% FBS conditions with minimal EGF-ligand present, the potency of afatinib on EGFR-KDD is approximately equivalent in the absence (0 g/ml) and presence (10 g/mL) of cetuximab (EC50 = 0.103 ± 0.035 nM and 0.095 ± 0.040 nM, respectively). Similar results are observed in Ex19Del (EC50 = 0.061 ± 0.027 nM and 0.060 ± 0.017 nM, respectively). The near complete ablation of Ba/F3 L858R viability at higher concentrations of cetuximab mask any potential similar effects. Generally, we observe that Ex19Del and L858R are more sensitive to afatinib than is EGFR-KDD (Figure 4.10B), consistent with our phosphorylation assays (Figure 4.10A and Figure 4.11A).

As the concentration of EGF-ligand in the medium is increased, we observe not only an increase in viability with cetuximab and increased EC50 of afatinib, but also a greater potentiation of afatinib by cetuximab (Figure 4.10B – C and Figure 4.11c – e). In 10% FBS + 50 ng/ml exogenous EGF, we observe a 5.8x increase in afatinib potency transitioning from 0 g/ml to 10 g/ml in Ba/F3 EGFR-KDD cells. We also observe potentiation of afatinib in Ex19Del (4.7x) and L858R (3.7x) (Figure 4.11E). Compared to Ex19Del and L858R, the larger potentiation of afatinib inhibition of Ba/F3 EGFR-KDD by cetuximab seems to be mediated by the lower inhibition of EGFR-KDD by afatinib. Together, our data suggests that a lower dose of afatinib can be administered to maximally inhibit EGFR-KDD when supplemented with cetuximab.

4.3 Discussion

In this study, we combined methods in clinical genomics, computational structural biology, biochemistry, and biophysics to mechanistically characterize a former VUS, EGFR exon 18–25 Kinase Domain Duplication (EGFR-KDD). To investigate the prevalence of KDD in all ERBB family members across various cancers, we analyzed comprehensive genomic profiling data from two large databases. We discovered that ERBB-KDDs are recurrent at a frequency between 0.58 - 2.4% in glioma, 0.07 - 0.22% in NSCLC, and 0.05 - 0.40%

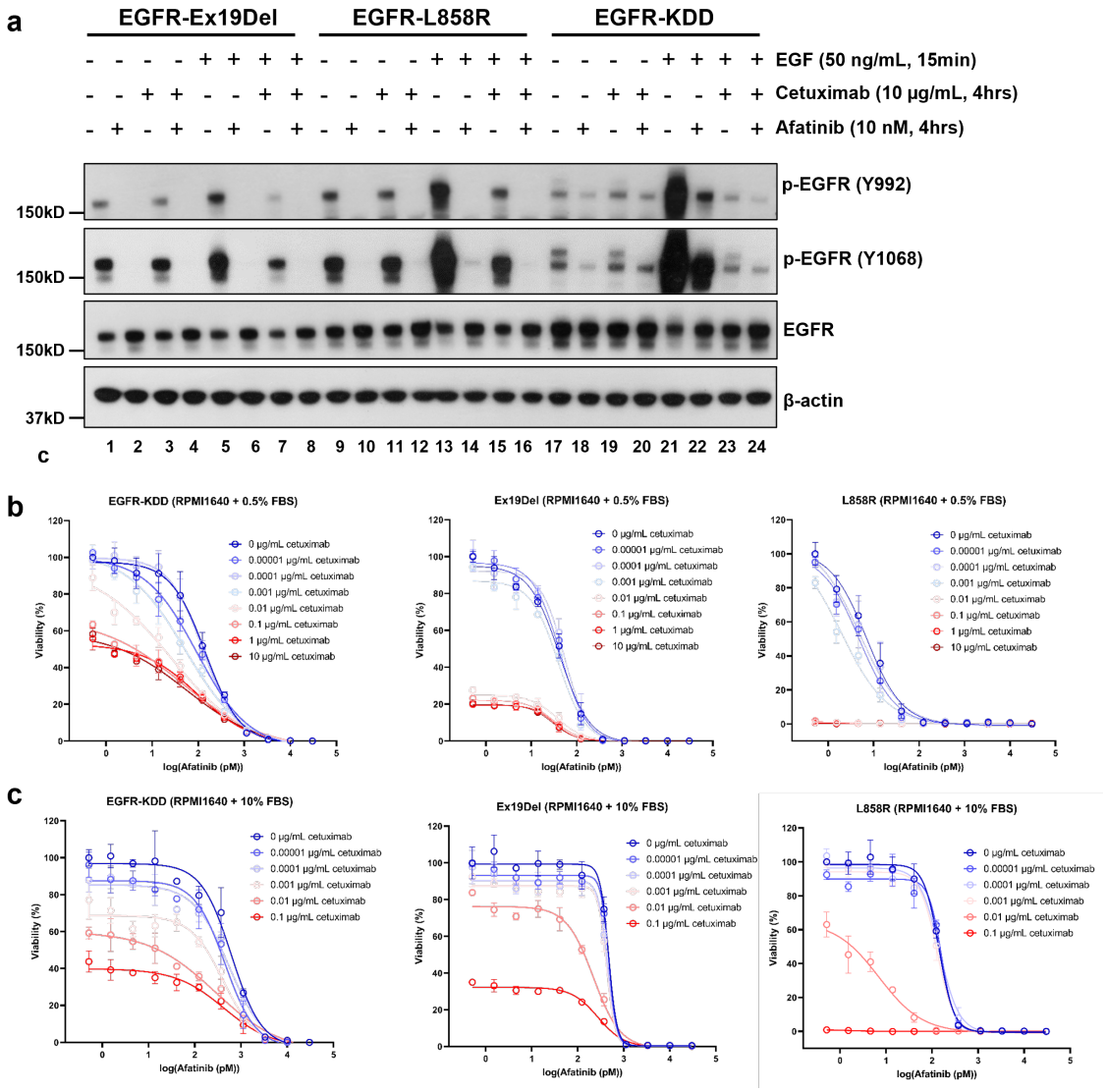


Figure 4.10: Inhibition of EGFR-KDD is maximally achieved by blocking both intra- and inter-molecular dimerization a, YAMC cells were starved for 12 hours and treated with afatinib (10 nM in serum-free medium) and cetuximab (10 μg/ml in serum-free medium) for 3 hours 45 minutes, and then were treated with EGF (50 ng/mL in serum-free medium) for 15 minutes. The cells were harvested and analyzed by Western blot. b, Cell Viability Assay was performed in mIL3-independent Ba/F3 cells stably expressing EGFR-KDD, Ex19Del and L858R supplemented with 0.5% FBS. 5,000 cells were seeded in 96-well plate with the treatment of afatinib and cetuximab. Three days after incubation, CellTiter-Blue Reagent was added, and the fluorescence was detected at 560EX/590EM with a Synergy HTX microplate reader (BioTek Instruments, Winooski, VT, USA). c, Cell Viability Assay was performed in mIL3-independent Ba/F3 cells stably expressing EGFR-KDD, Ex19Del and L858R supplemented with 10% FBS. For b and c, n=3 biologically independent samples were examined over 3 independent experiments. Data are presented as mean values +/- SD. Results in a, b and c are the representative of three independent experiments. Data and illustrations produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M.

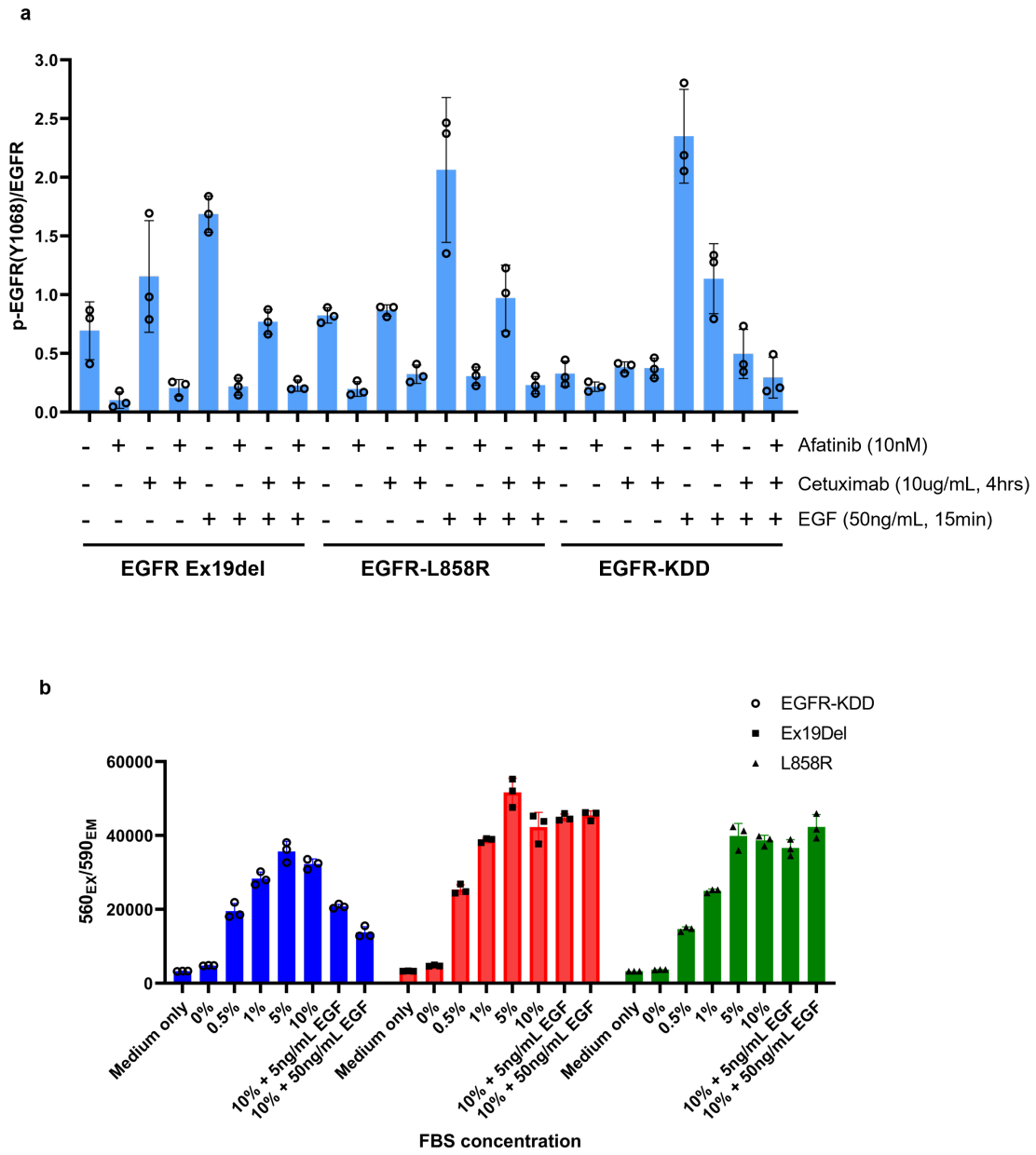


Figure 4.11: Inhibition of EGFR-KDD is maximally achieved by blocking both intra- and inter-molecular dimerization. a, Quantification of YAMC antibody/TKI treatment Western blots in Figure 4.10A. pEGFR/EGFR was presented as mean values of three independent experiments ± SD. b, BaF3 cell growth at different concentration of fetal bovine serum (FBS). 5,000 cells were seeded in 96-well plate with the treatment of afatinib and cetuximab. Three days after incubation, CellTiter-Blue Reagent was added, and the fluorescence was detected at 560EX/590EM with a Synergy HTX microplate reader (BioTek Instruments, Winooski, VT, USA). c, Cell Viability Assay was performed in mIL3-independent Ba/F3 cells stably expressing EGFR-KDD, Ex19Del and L858R in RPMI1640 supplemented with 10% FBS. d, Cell Viability Assay was performed in mIL3-independent Ba/F3 cells stably expressing EGFR-KDD, Ex19Del and L858R in RPMI1640 supplemented with 10% FBS and 5ng/mL EGF. e, Cell Viability Assay was performed in mIL3-independent Ba/F3 cells stably expressing EGFR-KDD, Ex19Del and L858R in RPMI1640 supplemented with 10% FBS and 50ng/mL EGF. Data and illustrations produced by Du, Z., Gallant, J.-N.; Zhang, Y.-K.; Yan, Y.; Red-Brewer, M., and Lovly, C. M.

in breast cancer. We identified fractions of KDDs in multiple other tumor types as well. No previous studies have reported KDD in ERBB2, ERBB3 and ERBB4. These data indicate that ERBB-KDDs account for a small but significant fraction of ERBB family-mediated cancers, and suggest utility of approved targeted therapies for patients based on standard of care clinical genomic testing. Importantly, developing targeted therapies for uncommon variants has precedent. ROS1 variants account for 1% of lung cancers49 and have been detected with lower prevalence in multiple other cancers50 and NTRK fusions have been implicated in 0.31% of adult tumors and in 0.34% of pediatric tumors51. There are TKIs targeting both ROS1 and NTRK52, 53 that are FDA approved and additional agents in clinical development. Further, in the case of KDD, both TKIs and antibody therapies already exist for ERBB receptors, thus new trials and therapeutic strategies for this population does not depend on new therapy development.

We sought to elucidate the mechanisms of EGFR-KDD-driven oncogenicity. We demonstrate that EGFR-KDD forms a catalytically active asymmetric intra-molecular dimer in the absence of EGF-ligand stimulation. Mutations disrupting the intra-molecular dimerization interface abolish the phosphorylation of EGFR-KDD in its monomeric form, and the loss of phosphorylation in these mutants can be recovered by the formation of inter-molecular dimerization and multimerization. These data demonstrate that ligand-independent constitutive activation EGFR-KDD is driven by asymmetric intra-molecular dimerization.

We next characterized differences in the functionality of the JMB region of EGFR-KDD relative to EGFR-WT. The JMB is a conserved stretch of amino acids critical for inter-molecular dimerization in wild-type HER-family receptor kinases. In EGFR-KDD, the JMB region of TKD2 is covalently linked to the C-terminus of TKD1. All-atom computational modeling investigations coupled with in vitro mutagenesis suggests the EGFR-KDD linker region is capable of forming specific stabilizing JMB domain contacts within the intra-molecular dimer; however, the forced proximity of the two kinase domains by the linker is sufficient for elevated EGFR-KDD activity relative to EGFR-WT. In comparison, EGFR-WT depends on stable contacts in the JMB domain for dimer activity (Jura et al., 2009; Red Brewer et al., 2009). We focused our analysis on EGFR-KDD with duplication of exons 18-25, but other groups have recently identified EGFR-KDD with longer duplications (e.g. exons 14-26 and exon 17-25) (Wang et al., 2019b) that may reduce the likelihood of forming stabilizing contacts at the linker JMB interface of the intra-molecular dimer. We speculate that there may be selective pressure for specific linker lengths/sequences in the formation of KDDs. Recent investigations have suggested similar structural constraints in the context of EGFR exon 19 deletion mutations (Foster et al., 2016).

EGFR-KDD further forms EGF-dependent inter-molecular dimers. Inter-molecular dimerization of EGFR-KDD increases activity in part by stabilizing the active conformation of the EGFR-KDD donor kinase domain. This has broad implications for HER-family signaling as well. We speculate that the formation of

dual activator/receiver kinases in higher order oligomers of HER-family receptors may contribute to ligand-dependent increases in phosphorylation29, 31. In the present study, we did not identify the configuration of the EGFR-KDD inter-molecular dimer/multimer. Mutations at N1 and C2 only partially disrupted EGF-dependent phosphorylation (Figure 4.6A, lane 7, 8, 17, 18; Figure 4.7A, lane 7, 8, 17, 18). Moreover, in the side-by-side structural model, the N-termini of TKD1 and TKD3 are oriented in close proximity (Figure 4.6C, yellow), while in the end-to-end model they are separated (Figure 4.6B, yellow). Consequently, we considered the end-to-end model less likely to form interactions between the N-terminal juxtamembrane A (JMA) and TM domains of the two interacting proteins, a key feature of inter-molecular dimerization in EGFR-WT (Jura et al., 2009; Red Brewer et al., 2009). Nevertheless, it is clear that EGFR-KDD is forming an EGF-dependent inter-molecular dimer. We anticipate that future investigations will identify the mostly likely inter-molecular configuration(s).

Interestingly, EGF-stimulated EGFR-KDD displays substantially more phosphorylation than EGF-stimulated canonical activating mutations (Figure 4.10A and Figure 4.7A). We speculate that this may be because of the increased ratio of EGF-ligand to active recruited kinase domains in EGFR-KDD (i.e. EGF-mediated dimerization of two extracellular domains results in an effective tetramer of intracellular kinase domains with potentially 2 – 3 active TKDs, versus typical oncogenic activation with 1 – 2 active TKDs). Alternatively, it may be that the EGFR-KDD inter-molecular dimer forms a more favorable interface than other oncogenic mutants, thus resulting in increased dimerization and activity. A combination of factors likely contributes to the overall increase in phosphorylation that we observe. Additional studies are needed to characterize the EGFR-KDD inter-molecular dimer.

Through a combination of biochemical and biophysical methods, we also determined that EGF-ligand stimulation induces formation of catalytically active homo- and hetero- inter-molecular dimers and multimers. Critically, this demonstrates that EGFR-KDD retains the ability to activate other ERBB family members. This has important implications for the therapeutic management of patients whose tumors harbor EGFR-KDD. Indeed, we found neither cetuximab nor afatinib alone were able to completely ablate EGFR-KDD phosphorylation. We demonstrate, however, that cetuximab can be used to potentiate afatinib inhibitory activity for greater overall inhibition. We suspect that this is because of the synergistic mechanisms of the two drugs: cetuximab disassembles dimers and removes the ability of EGFR-KDD to activate other ERBB kinases, and afatinib inhibits the active intra-molecular dimer EGFR-KDD. It has been well-recognized that cetuximab induces degradation of EGFR mutants in different NSCLC cells (Doody et al., 2007; Perez-Torres et al., 2006). In this study, no degradation of EGFR-Ex19Del, L858R and EGFR-KDD levels were observed in YAMC (Figure 4.10A) and NR6 cells (Figure 4.7B), probably due to the shorter treatment time than previous studies (4hrs versus 24 – 72hrs) (Doody et al., 2007; Perez-Torres et al., 2006).

Finally, our computational and biochemical insights raise important considerations for the use of EGFR-KDD as a research tool. Whereas the inactive form of EGFR can be readily studied by the introduction of inter-molecular dimer-disrupting interface mutations (Zhang et al., 2006), controlling the active fraction of EGFR in vitro has typically required introduction of known oncogenic point mutations or stimulation with EGF-ligand. The former causes well-documented perturbations to enzyme kinetics (Yun et al., 2007; Carey et al., 2006; Gilmer et al., 2008; Yun et al., 2008a), while recent literature has demonstrated that the latter can influence EGFR multimerization and phosphorylation in a concentration-dependent manner (Needham et al., 2016a). Moreover, dimerization and activation of EGFR oncogenic missense mutants is dependent on protein concentration (Sholl et al., 2009) and/or EGF- ligand stimulation (Red Brewer et al., 2013). EGFR-KDD provides a model of a fully active EGFR dimer in an EGF-independent setting, and may provide a more native-like control than kinase domain missense mutants without the complexity of concentration-dependent signaling effects.

Kinase Domain Duplications (KDDs) represent a novel form of activation for oncogenic kinases via a mechanism of constitutive dimerization. In this study, we have systematically characterized the fundamental biochemical and biophysical features of a prototypical KDD, EGFR-KDD. Subsequently, we identified potential treatment strategies in pre-clinical models of EGFR-KDD-mediated disease. This represents the first comprehensive mechanistic and pre-clinical evaluation of treatment strategies specifically for a KDD-mediated disease. We anticipate that our results will also be used to inform additional studies on kinase duplication domains.

4.4 Methods

4.4.1 Cell Culture, Reagents and Transfection

Ba/F3 cells were purchased from DSMZ. NR6 cells were a kind gift from Dr. William Pao (Regales et al., 2009). YAMC EGFR-/- cells were a kind gift from Dr. Robert H. Whitehead (Dise et al., 2008). Plat-GP cells were purchased from CellBioLabs. HEK293 cells were purchased from ATCC. Ba/F3 cells were maintained in RPMI 1640 medium (Mediatech, Inc.) supplemented with 1 ng/mL murine IL3 (Gibco, Life Technologies). NR6 cells were maintained in DMEM (Gibco). The Plat-GP cell line was cultured in full DMEM with selection of 1 g/mL blasticidin (Gibco). YAMC cells were cultured as previously described (Dise et al., 2008; Whitehead et al., 1993), 64. COS-7 cells were cultured in DMEM (Calsson Lab, Smithfield, UT). All media were supplemented with 10% heat inactivated FBS (Gibco) and penicillin-streptomycin (Gibco) to final concentrations of 100 U/mL and 100 g/mL, respectively. All cell lines were maintained in a humidified incubator with 5% CO2 at 37°C (33°C for YAMC cells (Whitehead et al., 1993)) and routinely evaluated for mycoplasma contamination.

Cetuximab was purchased from Bristol-Myers Squibb (Princeton, NJ). mAb806 is produced and purified in the Biological Production Facility (Ludwig Institute for Cancer Research, Melbourne) (Johns et al., 2003, 2002). Transient transfection for expression in HEK293 cells was carried out using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. A total of 0.45 g of each expression plasmid was used per well in 6 well plates. To assess ligand-dependent EGFR activation, cells were serum starved overnight and treated with 50 ng/mL EGF for 5min.

For PIE-FCCS experiments, COS-7 cells were transiently transfected 24 hours before the experiment using Lipofectamine 2000 (Invitrogen). A total of 5 µg DNA (1:1 ratio of mCherry-tagged and eGFP-tagged plasmids mixture) was used per 35 mm MatTek plate (MatTek Corporation, Ashland, MA) to express both fluorescent-tagged species evenly and acquire the local density of 100-2000 receptors/µm2. The media was changed to Opti-MEM I Reduced Serum Medium without phenol red (Thermo Fisher Scientific) before placing the plate in the on-stage incubator (37 °C) for FCCS measurement. Measurements were taken for both ligand-free and ligand-stimulated state of each construct, with 2 g/mL recombinant human EGF (Sigma Aldrich, St. Louis, MO) or NRG1 (RD Systems, Inc., Minneapolis, MI) as the ligand.

4.4.2 Plasmid Construction

Generation of EGFR-KDD, EGFR-WT and EGFR-L858R constructs was described previously7. EGFR-KDD mutations were constructed by using multisite-directed mutagenesis (Agilent) on the pMa-EGFR-KDD plasmid per the manufacturer's recommendations - with the exception of extension time being set at 1.5 mins / kb. To specifically introduce mutations into each TKD due to the presence of two identical TKDs at the genomic level, after bi-directional dideoxy sequencing, pMa-EGFR-KDD-mutants were digested with ClaI and recombined with other pMa-EGFR-KDD fragments to create all single mutants: ClaI digests mutated pMa-EGFR-KDD plasmid were recombined with a ClaI–ClaI segments from unmutated pMa-EGFR-KDD and/or ClaI digests of unmutated pMa-EGFR-KDD plasmid were recombined with ClaI–ClaI segments from mutated pMa-EGFR-KDD. pMa-EGFR-KDD mutants were then subcloned to the pMSCV vector by HpaI digest and then subcloned to pcDNA3.1(-) vector by XhoI /HindIII digest. All plasmids were verified in the forward and reverse directions by Sanger sequencing. To obtain V5-epitope tagged EGFR-KDD, we used PCR to add AgeI to the 3' end of EGFR-KDD fragment by using pMSCV-EGFR-KDD as template, then EGFR-KDD fragment was inserted into pcDNA6-V5 HisB vector by using SnaBI and XhoI. To obtain Myc-epitope tagged EGFR-KDD, the EGFR-KDD fragment was subcloned to pEF4Myc-HisB vector by using MfeI and XhoI. pcDNA6-EGFR-WT with Myc-epitope tag was purchased from Addgene (42665). V5-epitope tagged HER2 and HER3 were kind gift from Dr. Carlos L. Arteaga67. For PIE-FCCS experiments, EGFR-WT, HER2 and HER3 was subcloned to eGFP-N1 and mCherry-N1 vectors by XhoI and AgeI digest.

EGFR-KDD was subcloned to eGFP-N2 and mCherry-N2 vectors by using SnaBI and XhoI digest. For V5 tagged epitope EGFR-WT, we replaced the eGFP fragment of pEGFR-N1-EGFR-WT by V5 tagged epitope. In this study, for EGFR mutations, we utilized codon numbering of the human immature EGFR sequence that includes the 24-residue signal sequence.

4.4.3 Generation of stable cell lines

Constructs of pMSCV, EGFR-WT, EGFR-L858R, EGFR-KDD and EGFR-KDD-I706Q, D837N, V948R, I1057Q, D1188N and V1299R mutations were introduced into NR6 and YAMC cells separately by retroviral transduction system as described previously7. Construct of EGFR Ex19Del (E746_A750del) was stably introduced into YAMC cells, and constructs of EGFR Ex19Del (E746_A750del), EGFR-L858R and EGFR-KDD were stably introduced into Ba/F3 cells as described previously (Brown et al., 2019a).

4.4.4 Immunoblotting and Antibodies

For immunoblotting, cells were washed in cold PBS, and lysed in RIPA buffer (150 mmol/L NaCl, 1% Triton-X-100, 0.5% Na-deoxycholate, 0.1% SDS, 50 mmol/L Tris-HCl, pH 8.0) with freshly added 40 mmol/L NaF, 1 mmol/L Na3VO4, and protease inhibitor (Thermo Fisher Scientific, Waltham, MA). Lysates were quantified by Bradford assay in SmartSpec Plus Spectrophotometer (Bio-Rad, Hercules, CA) following the manufacturer's instructions. Lysates were subjected to SDS-PAGE followed by blotting with the indicated antibodies and detection by Western Lightning ECL reagent (Perkin Elmer, Waltham, MA). The densitometry for both phosphotyrosine content at Y1068 and total EGFR expression was quantified by ImageJ Software. The ratio of phosphotyrosine to total EGFR expression for each construct relative to EGFR-KDD was calculated. All immunoblotting experiments were performed three independent times and one representative replicate was shown in the manuscript. Raw uncropped and unprocessed scans of all blots, quantifications and standard deviations were included in the Source Data file.

For co-immunoprecipitation experiments, cells were washed in cold PBS and lysed in hypotonic buffer (20mM HEPES pH7.5, 10mM KCl, 1mM EDTA, 1mM EGTA, 1mM mgCl2, 0.1% NP-40, EDTA-free Protease Inhibitor Cocktail (Sigma-Aldrich 04693159001)). The lysates were supplemented with 150 mM NaCl before centrifuging. Protein G Dynabeads (10004D, Life Technologies, Carlsbad, CA) were incubated with the primary antibody for 30 minutes at room temperature. Lysates were then added and incubated for 3 hours at 4°C. Immobilized beads were washed three times with hypotonic buffer supplemented with 0.65 M NaCl. 2xSDS loading buffer was added to the beads and then used for immunoblotting analysis. All co-immunoprecipitation experiments were performed two independent times and one representative replicate was shown in the manuscript.

4.4.5 Antibodies

Antibodies used included: EGFR (1:2000, 4267), phospho-EGFR (Y992) (1:1000, 2235), phospho-EGFR (Y1068) (1:1000, 2234), phospho-EGFR (Y1173) (1:1000, 4407) (For EGFR phosphorylation sites, we utilized codon numbering of mature EGFR sequence that does not include the 24-residue signal sequence), horseradish peroxidase (HRP) - conjugated anti-mouse (1:5000, 7076), and HRP-conjugated anti-rabbit (1:5000, 7074) (Cell Signaling, Beverly, MA); V5 (1:5000, MCA1360GA, AbD Serotec), Myc (1:2500, Sigma-Aldrich A5963); actin antibody (1:5000, Sigma-Aldrich A2066).

4.4.6 Pulsed Interleaved Excitation Fluorescence Cross-Correlation Spectroscopy (PIE-FCCS)

FCCS data were taken on a customized inverted microscope setup coupled with pulsed interleaved excitation and time-correlated single photon detection as described in previous works (Huang et al., 2016; Endres et al., 2013). A supercontinuum pulsed laser (9.2 MHz repetition rate, SuperK NKT Photonics, Birkerød, Denmark) was split into two beams of 488 nm and 561 nm through a series of filters and mirrors for the excitation of eGFP and mCherry respectively. The beams were directed through two different-length single mode optical fiber to introduce 50 ns time delay for pulsed interleaved excitation to eliminate possible spectral crosstalk (Comar et al., 2014). The beams were overlapped before entering the microscope through a dichroic beam splitter (LM01-503-25, Semrock) and a customized filter block (zt488/561rpc, zet488/561m, Chroma Technology). A 100X TIRF oil objective (Nikon, Tokyo, Japan) was used for the excitation beam focus and fluorescence emission collection. A short fluorescently tagged DNA fragment was used to verify the alignment of the system, including the confocal volume overlap. Negative and/or positive controls were tested regularly prior to the experimental samples for comparisons of the fit parameters. The excitation beams were focused to the peripheral membrane of the cell to allow the fluorescence measurements of only the membrane-bound receptors. Data were only taken on the flat, peripheral membrane area, where the distance between the basal and apical membranes were within a few hundred nanometers, to avoid inclusion of fluorescence from cytosolic organelles or vesicles. For each cell, one area of the membrane was selected for data collection. Six 10-second acquisitions were taken per area. The fluorescence signal was collected through a home-built confocal detection unit with a 50 μm confocal pinhole and dichroic beam splitter (LM01-503-25, Semrock, Rochester, NY). The two signals were filtered (91032, Chroma Technology Corp., Bellows Falls, VT; zt488/561rpc and zet488/561m, Chroma Technology Corp., Bellows Falls, VT) and then focused independently on to single-photon avalanche diodes (Micro Photon Devices, Bolzano, Italy). The photon counts were recorded by a time-correlated single photon counting module (Picoharp 300, PicoQuant, Berlin, Germany). For analysis, the time-tagged photon data were gated to isolate photons that arrived within 40 ns after each laser pulse arrival time. Then we calculated auto- and cross-correlation curves correspond-

ing to each species using our custom MATLAB script. Curves of six consecutive acquisitions per area were averaged then fitted to a single component, 2D diffusion model as described in previous works (Endres et al., 2013; Kaliszewski et al., 2018; Comar et al., 2014).

The auto-correlation curves contain two types of decay. The first decay is due to the photophysical activity, such as triplet relaxation or blinking. The second decay indicates the average dwell time (τD), which is used to calculate the effective diffusion coefficient using Deff = ωo2/4τD. The amplitude of the correlation curves indicates local concentration of the diffusing receptors. Using the cross-correlation curve, we can calculate cross-correlation values, or fraction correlated (fc) values that indicate the degree of oligomerization. For an ideal system undergoing on dimerization, the fc value varies from 0 to 1, with 0 indicating the system is monomeric and 1 indicating complete dimerization. For real systems, effects like photostability, interaction statistics, and relative expression levels drop the expected fc value for dimerization into the range of 0.10 to 0.15 for a monomer-dimer equilibrium. For higher order oligomerization the fc values will increase, allowing us to compare the degree of oligomerization for more complex systems (Comar et al., 2014).

4.4.7 Anchorage-Independent Assays and Cell Viability Assay

Anchorage-independent assays were performed as previously described (Borowicz et al., 2014; Horibata et al., 2015). For the bottom layer of agar, 1.5 mL of a 1:1 mix of 1.0% agar (prepared in 1xPBS) and medium was plated in each well of 6-well plate. For the upper layer of agar, 1.5 mL of a 1:1 mix of 0.6% agar (prepared in 1xPBS) and medium containing 5,000 cells was plated into each well of 6-well plate. Colonies were counted using GelCount (Oxford Optronix) with identical acquisition and analysis settings. Cell viability assay was performed on IL3 independent Ba/F3 cells stably expressing EGFR-KDD, Ex19Del and L858R by using CellTiter-Blue® Cell Viability Assay (G8080, Promega, Madison, WI) following manufacturer's instructions. All experiments of anchorage-independent assays and cell viability assay were performed three independent times in triplicate, and one representative replicate was shown in the manuscript.

4.4.8 Molecular Modeling

Previously, we performed de novo loop modeling to determine a geometrically plausible model of the EGFR-KDD linker region (Gallant et al., 2015). Here, an all-atom structural model of the EGFR-KDD intracellular domain was generated with RosettaCM (Song et al., 2013) with the active EGFR WT dimer PDB ID 2GS6 as the base template. Missing density in the β3-αC region was templated with PDB ID 2ITX. The N- and C-termini of the donor and receiver kinases of the EGFR-KDD intra-molecular dimer, respectively, as well as the connecting linker region, are based on three templates: the previously modeled linker region from Gallant et al. 2015 (Gallant et al., 2015); the JMB domain of PDB ID 4RIW; and the JMB domain of PDB ID 3GOP.

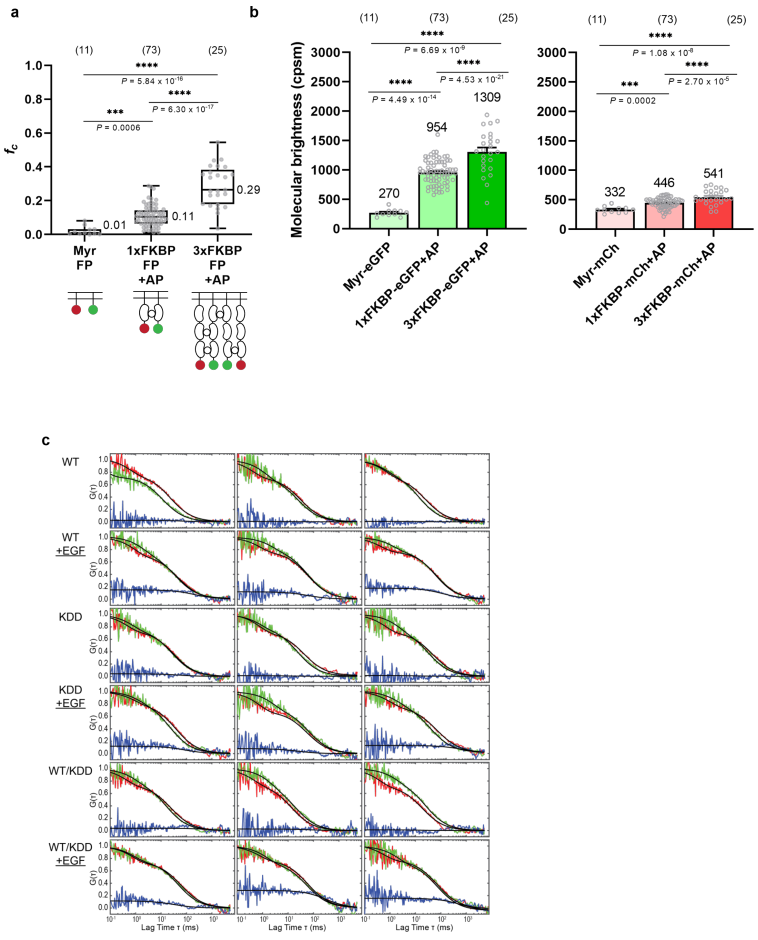


Figure 4.12: EGF ligand stimulation induces the formation of EGFR-KDD inter-molecular dimers. a, Cross correlation values of PIE-FCCS control constructs. The monomer control (Myr-FP: myristoylated fluorescent protein [mCh or eGFP; coexpressed together]) had an fc value of 0.01 indicating no interaction. Upon cross-linking by a synthetic dimerizer (AP: AP20187) the dimer control (1xFKBP-FP) had an average fc value of 0.11, consistent with dimerization. The multimer control (3xFKBP-FP) had an fc value of 0.29 consistent with the formation of a mixture trimer and tetramer species. b, Average molecular brightness of PIE-FCCS negative and positive controls in Figure 4.7c (Left: constructs with eGFP tag; right: constructs with mCh tag). The oligomer control (3xFKBP+AP) has much higher molecular brightness as expected due to clustering. mCh-tagged constructs show subtle changes in the molecular brightness due to the photophysical properties of mCherry. However, the molecular brightness changes are still statistically significant between all constructs. c, Representative FCCS data for EGFR-WT and EGFR-KDD expressed in COS-7 cells. The scatter plot connected with red, green and blue lines indicates the normalized auto-correlation function for mCherry-fused/eGFP-fused receptors and cross-correlation function, respectively. Black solid line shows the fit model of each curves. For a and b, the numbers in parenthesis above the boxplot/bar graph are the total number of cells where data were taken on. Both One-Way ANOVA test and Uncorrected Fisher's LSD test were down to obtain adjusted and individual p values.

Missing residues are modeled de novo with RosettaCM fragment insertion. Three rounds of comparative modeling were performed. After rounds two and three, the best scoring models with varying RMSDs from the lowest scoring model in each round were selected as additional starting templates for the next round. After the third round, distance-based clustering of the linker region identified three low energy clusters. The best scoring model from each cluster was refined with a 1 s molecular dynamics (MD) simulation in Amber18 (Case et al., 2018). The final EGFR-KDD model and EGFR-WT homodimer subsequently each underwent 1 μs MD simulations.

Models were solvated in a rectangular box of SPC/E explicit solvent neutralized with monovalent anions. Protein was buffered on all sides with 12 $\AA$ solvent. Solvent and ions were minimized with 500 steps steepest gradient descent followed by 1000 steps of conjugate gradient descent while protein atoms were restrained with a force constant of 10.0 kcal/mol/$\AA^2$. The protein was then minimized for 200 steps steepest gradient descent followed by 800 steps of conjugate gradient descent in buffer restrained with a force constant of 5.0 kcal/mol/A2. Finally, restraints were removed from the system for 100 additional steps of steepest gradient descent followed by 900 steps of conjugate gradient descent minimization.

Post-minimization, SHAKE was implemented to constrain covalent bonds to hydrogen atoms. Systems were slowly heated in NVT ensemble to 100K over 50 ps with a 1 fs timestep. Subsequently, systems were heated in NPT ensemble at 1 bar with isotropic position scaling from 100K to 300K over 500 ps and 1 fs timestep. Equilibration/production simulations were run in the NPT ensemble at 300K with a Monte Carlo barostat. Temperature was controlled using Langevin dynamics with a collision frequency of 1 ps^{-1} and a unique random seed for each simulation. Periodic boundary conditions were imposed on the system throughout heating and equilibration. Electrostatics were evaluated using the Particle Mesh Ewald (PME) method and a distance cutoff of 8.0 $\AA$. A 2 fs integration timestep was employed during production simulations. All RMSD and RMSF calculations were performed with CPPTRAJ (Roe and Cheatham, 2013).

Approximations of the linker interaction energies of the top three EGFR-KDD clusters were performed with the single-trajectory molecular mechanics / generalized Born solvent-accessible surface area (MM-GBSA) method as implemented in MMPBSA.py (Miller et al., 2012). GBSA was calculated with the OBCII Generalized born solvent model with a surface tension of 0.0072 kcal/mol/$\AA$2 and salt concentration of 0.15 M, and nonpolar contributions to the solvation free energy were computed with the LCPO method. Entropic contributions to binding were neglected. The final reported values are averaged over frames collected every 100 ps.

4.4.9 Kinase Domain Duplication Detection from Foundation Medicine and MSK-IMPACT datasets

For the Foundation Medicine dataset, a minimum of 50 ng of DNA was extracted from formalin-fixed paraffin-embedded sections and comprehensive genomic profiling was performed on hybridization-captured, adaptor ligation-based libraries to a median exon coverage depth of ¿ 500X for all coding exons of 315 (FoundationOne®, n = 152,674), or 324 (FoundationOneCDx®, n = 86,824) cancer-related genes plus selected introns from genes frequently rearranged in cancer to identify base substitutions, small insertions or deletions, copy number alterations (focal amplifications and homozygous deletions), and rearrangements, as previously described (Frampton et al., 2013). Testing was performed in a Clinical Laboratory Improvement Amendments-certified, College of American Pathologists-accredited reference laboratory (Foundation Medicine, Cambridge, MA). We interrogated the Foundation Medicine dataset of n = 239,498 consecutive unique solid tumor specimens for kinase domain duplications (KDD) in EGFR, ERBB2, ERBB3 and ERBB4. These rearrangement duplications were detected by clustering chimeric and semi-mapped paired-end reads within each gene of interest and mapping breakpoints onto the hg19 reference genome assembly, as previously described75. A KDD was therein defined as a large genomic duplication where breakpoints both flanked and did not disrupt the region corresponding to the respective gene's kinase domain. Statistical enrichment including p-value and odds-ratio (OR) was calculated using Fisher's exact testing. For Foundation Medicine cases, approval for this study, including a waiver of informed consent and a Health Insurance Portability and Accountability Act waiver of authorization, was obtained from the Western Institutional Review Board (protocol no. 20152817).

MSK-IMPACT sequencing data from patients whose tumor and matched normal samples were prospectively sequenced between January 2014 and September 2019 (n=40,165) were used in this study. Structural variant detection was performed on the paired-end reads using Delly (version 0.7.5; https://github.com/dellytools/delly). Duplication events that surrounded or overlapped known kinase domains were selected for further manual review. For copy number-based analysis, coverage data from the tumor and an unmatched normal sample were used to generate a fold change value (Ross et al., 2017) for each exon in a kinase gene. Using k-mean clustering (k = 2), we identified samples where one of the clusters was overlapping (requiring at least 70% of the kinase domain to be involved) or encompassing the kinase domain with a median cluster fold change difference of at least 0.4. We combined the two datasets for further manual review to identify a subset of confident KDD calls.

4.4.10 Statistical analysis

Statistical significance was analyzed using unpaired Student's t-test for two groups or one-way ANOVA for multiple groups. Results were displayed as mean values ± standard deviation (SD). For all tests, the criteria

for significance were P ¡ 0.05 (*), P ¡ 0.01 (**), and P ¡ 0.001 (***). Statistical analysis was carried out using Prism 9 (GraphPad Software).

CHAPTER 5

Co-Occurring Gain-of-Function Mutations in HER2 and HER3 Modulate HER2/HER3 Activation, Oncogenesis, and HER2 Inhibitor Sensitivity

This chapter is taken from Hanker, A. B.*; Brown, B. P.*; Meiler, J.*; Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; Sheehan, J. H.; He, J.; Lalani, A. S.; Arteaga, C. L. Cancer Cell 2021, 39 (8), 1099-1114.e837 (*These authors contributed equally).

5.1 Introduction

Activating mutations in HER2 (also known as ERBB2) are oncogenic drivers in a subset of breast and other cancers (Bose et al., 2013; Hanker et al., 2017; Hyman et al., 2018). In breast cancer, HER2 mutations typically occur in the absence of HER2 amplification, are more common in invasive lobular breast cancer (Deniziaut et al., 2016; Desmedt et al., 2016; Ping et al., 2016; Ross et al., 2013), and are associated with poor prognosis (Kurozumi et al., 2020; Ping et al., 2016; Wang et al., 2017). Recurrent HER2 mutations promote resistance to antiestrogen therapy in estrogen receptor-positive (ER+) breast cancers (Croessmann et al., 2019; Nayar et al., 2019) and are found in $\tilde{5}\%$ of endocrine-resistant metastatic breast cancers (Razavi et al., 2018). They have also been implicated in resistance to HER2 inhibitors in HER2-amplified breast cancers (Cocco et al., 2018; Xu et al., 2017) and can be targeted with HER2 tyrosine kinase inhibitors (TKIs), such as neratinib. Approximately 30% of HER2-mutant metastatic breast cancers respond to neratinib (Hyman et al., 2017, 2018), suggesting that co-occurring mutations may modulate HER2 TKI response.

HER2 is a member of the ERBB receptor tyrosine kinase family, which includes EGFR, HER3 (ERBB3), and HER4 (ERBB4). Upon ligand-induced homo- and heterodimerization of the extracellular domain (ECD), ERBB receptors undergo a conformational change that triggers asymmetric dimerization of the kinase domains (KDs), leading to kinase activation and subsequent signal transduction through oncogenic pathways, such as the phosphoinositide-3-kinase (PI3K)/AKT/mTOR and RAS/RAF/MEK/ERK pathways (Zhang et al., 2006). Although HER2 lacks a high-affinity ligand, its natural conformation resembles a ligand-activated state and is the preferred heterodimer of EGFR and HER3 (Arteaga and Engelman, 2014). HER3 is catalytically impaired and its signaling depends on heterodimerization with catalytically active partner, such as EGFR and HER2 (Wallasch et al., 1995).

The most common HER2 mutations in breast cancer are missense mutations in the KDs, such as $HER2^{L755S}$ and $HER2^{V777L}$. While HER2 missense mutants exhibit gain-of-function activity (Bose et al., 2013), they are

not potently transforming in the absence of amplification and may require cooperation with other oncogenes to confer a fully transformed phenotype. For example, co-occurring PIK3CA mutations (encoding PI3K) cooperate with HER2 mutations to augment pathway activation (Zabransky et al., 2015). However, PIK3CA mutations are only found in Ĩ/3 of HER2-mutant breast cancers; other alterations that cooperate with HER2 mutations are not known.

Gain-of-function mutations in HER3 are found in 2% of breast cancers (Cancer Genome Atlas, 2012; Jaiswal et al., 2013; **?**). HER2/HER3 heterodimers exhibit high catalytic activity, strongly activate the PI3K/AKT/mTOR pathway, and induce transformation more potently than any other ERBB dimers (Choi et al., 2020; Holbro et al., 2003; Yarden and Sliwkowski, 2001). In the HER2/HER3 asymmetric dimer, the HER3 KD serves as the "activator," stimulating the kinase activity of the HER2 "receiver" (Choi et al., 2020). Co-occurring HER3 mutations have previously been found in HER2-mutant tumors (Hanker et al., 2017) and are associated with lower clinical response to neratinib in the clinic (Hyman et al., 2018; Smyth et al., 2020). We hypothesized that the mutant HER3 receptor cooperates with mutant HER2 to promote tumor growth via enhanced HER2 and PI3K activation.

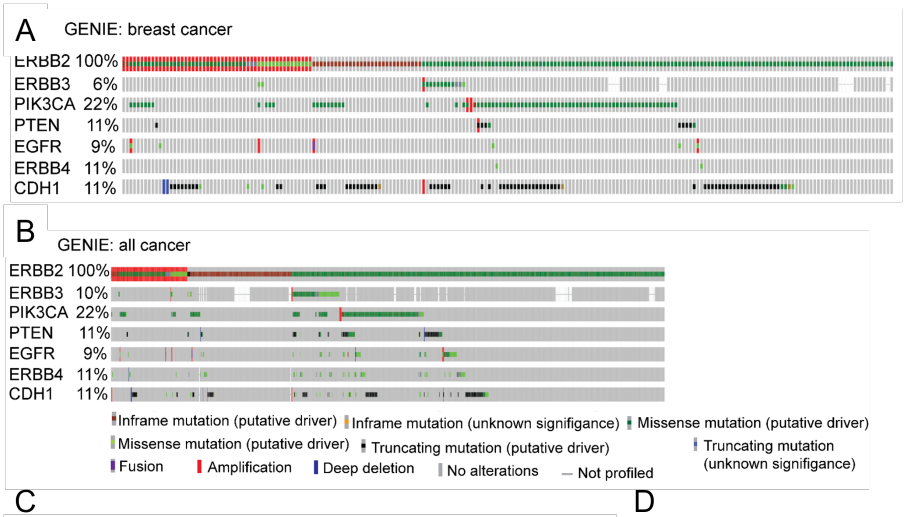
5.2 Results

5.2.1 Activating mutations in HER2 and HER3 co-occur in breast and other cancers

We interrogated 277 breast cancers (Figures 5.1A and 5.2A) and 1,561 pan-cancers harboring somatic HER2 mutations from the Project GENIE dataset (genie.cBioPortal.org) for co-occurring alterations in EGFR, ERBB3, ERBB4, PIK3CA, and PTEN (Figures 5.1B and 5.2B). Since HER2 mutations are known to be associated with lobular breast cancer (Desmedt et al., 2016), we also included the CDH1 gene, which is mutated frequently in lobular breast cancer. Mutations in HER2 and HER3 showed a significant tendency to co-occur in breast cancer (q = 0.006) and in all cancers ($q = 1.01 \times 10^{-26}$; Figures 5.1C and 5.2C).

Most co-occurrences were between known activating missense mutations in both genes rather than variants of unknown significance (Figures 5.2A and 5.2B). In breast cancer, neither EGFR nor ERBB4 alterations were found to co-occur with HER2 (Figure 5.2C). We also noted that HER3 mutations did not co-occur with HER2 in-frame insertion mutations or when HER2 was both mutated and amplified (Figures 5.1A and 5.1B). Intriguingly, in HER2-mutant breast cancers, co-occurring HER3 mutations were mutually exclusive with co-occurring PIK3CA, suggesting that HER3 and PIK3CA mutations are functionally redundant.

To identify the most common co-occurring HER2 and HER3 mutant allele pairs in breast cancer, we expanded our search to include additional datasets from Foundation Medicine and cBioPortal. We identified 67 breast cancers harboring mutations in both genes. The most common HER2 mutations were L755S (n = 24), S310F/Y (n = 16), V777L (n = 14), and L869R/Q (n = 7). The most common HER3 mutations



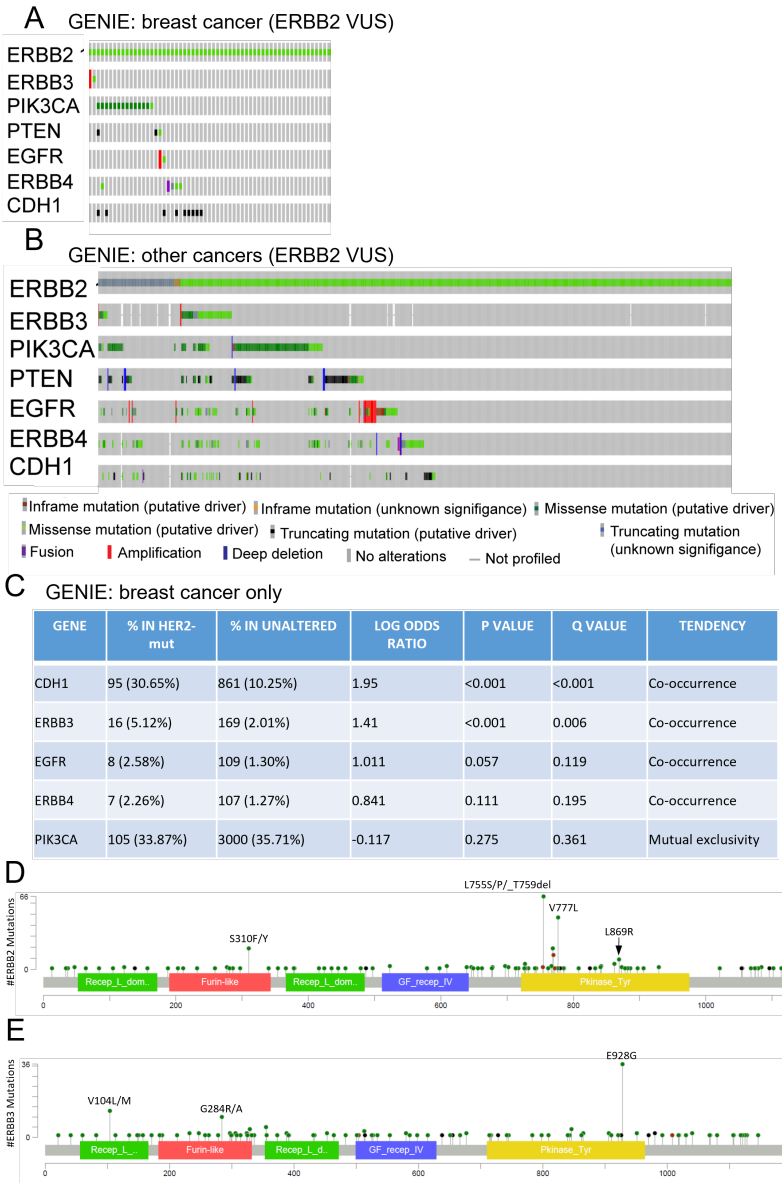
C

GENE	% IN HER2-mut	% IN UNALTERED	LOG ODDS RATIO	P VALUE	Q VALUE	TENDENCY
ERBB3	90 (8.73%)	774 (2.01%)	2.12	4.50E-29	1.01E-26	Co-occurrence
CDH1	97 (9.41%)	922 (2.39%)	1.98	3.26E-28	6.56E-26	Co-occurrence
ERBB4	104 (10.09%)	1090 (2.83%)	1.84	4.53E-27	7.58E-25	Co-occurrence
PIK3CA	188 (18.23%)	4428 (11.48%)	0.67	2.68E-10	3.77E-09	Co-occurrence
EGFR	88 (8.54%)	2094 (5.43%)	0.65	3.53E-05	2.29E-04	Co-occurrence

D

HER2 variant	HER3 variant	No. of co-occurrences (n=67 breast)
L755S	E928G	14
V777L	E928G	10
L869R/Q	E928G	5
S310F/Y	E928G	4
S310F/Y	V104L/M	4

Figure 5.1: ERBB2 and ERBB3 mutations co-occur in breast and other cancers. (A) 277 breast cancers with ERBB2 mutations and (B) 1,561 ERBB2-mutant cancers (all tumor types) in the Project GENIE database were interrogated for co-occurring alterations in the indicated genes. ERBB2 variants of unknown significance (VUS) are excluded. (C) Mutations in the indicated genes were analyzed for co-occurrence or mutual exclusivity with ERBB2 mutations using cBioPortal. (D) The most common co-occurring HER2/HER3 mutations in breast cancer were determined using databases from Project GENIE, cBioPortal [TCGA, METABRIC, MBC Project, Mutational Profiles of MBC (France), and Breast Invasive Carcinoma (Broad)], and Foundation Medicine. Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.



GENE	% IN HER2-mut	% IN UNALTERED	LOG ODDS RATIO	P VALUE	Q VALUE	TENDENCY
CDH1	95 (30.65%)	861 (10.25%)	1.95	<0.001	<0.001	Co-occurrence
ERBB3	16 (5.12%)	169 (2.01%)	1.41	<0.001	0.006	Co-occurrence
EGFR	8 (2.58%)	109 (1.30%)	1.011	0.057	0.119	Co-occurrence
ERBB4	7 (2.26%)	107 (1.27%)	0.841	0.111	0.195	Co-occurrence
PIK3CA	105 (33.87%)	3000 (35.71%)	-0.117	0.275	0.361	Mutual exclusivity

Figure 5.2: Gain-of-function, but not passenger, missense mutations in ERBB2 and ERBB3 have a tendency to co-occur. (A) Breast cancers and (B) all cancers with ERBB2 VUS in the Project GENIE database were interrogated for co-occurring alterations in the indicated genes. (C) Mutations in the indicated genes were analyzed for co-occurrence or mutual exclusivity with ERBB2 mutations in breast cancers from Project GENIE using cBioPortal. (D,E) Lollipop plots of ERBB2 (D) and ERBB3 (E) mutations in breast cancer from Project GENIE. Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

were E928G (n = 35), V104L/M (n = 8), T355A/I (n = 5), and K329E/I (n = 5). These were similar to the most common single HER2 and HER3 missense mutations found in breast tumors (Figures 5.2D and 5.2E). The most common pairs are shown in Figure 5.2D. Since $HER3^{E928G}$ is the most common co-mutated HER3 allele, we focused our studies on that mutation paired with $HER2^{L755S}$, $HER2^{V777L}$, $HER2^{L869R}$, and $HER2^{S310F}$.

5.2.2 Co-occurring HER2/HER3 mutants enhance KD dimerization and HER2 kinase activation

To determine the mechanisms of activation of mutant HER2 and HER3, we systematically evaluated the contributions of each mutation to HER2 kinase activation and HER2/HER3 dimerization (Figures 5.4A–5.4C). Previous work demonstrated an increase in $HER2^{WT}$ kinase activity when bound to $HER3^{E928G}$ relative to $HER3^{WT}$ (Collier et al., 2013). Subsequent work showed that $HER3^{E928G}$ enhances EGFR/HER3 dimerization affinity, potentially as a result of charge neutralization at the asymmetric dimer interface. However, neutralization of a glutamate interface residue in EGFR resulted in <2-fold increase in dimerization affinity, suggesting that charge neutralization may not be the primary contributor to $HER3^{E928G}$ gain of function (Littlefield et al., 2014). Therefore, we probed the effects of $HER3^{E928G}$ on HER2/HER3 dimerization using a combination of Rosetta DDG calculations and molecular dynamics (MD) simulations.

Consistent with previous studies, our Rosetta simulations suggest an enhanced dimerization affinity of $HER2^{WT}$/$HER3^{E928G}$ relative to $HER2^{WT}$/$HER3^{WT}$ (Figure 5.3A). Per-residue decomposition of Rosetta binding energy suggests that the largest contributions can be attributed to HER2 L790 and HER3 G927 (Figures 5.3B, 5.4D, and 5.4E). MD simulations displayed a reduced HER2 L790-HER3 G927 backbone hydrogen bond (H bond) distance (Figures 5.3C and 5.3D) and a $\tilde{1}$.3 kcal/mol increase in H bond stability in $HER2^{WT}$/$HER3^{E928G}$ relative to $HER2^{WT}$/$HER3^{WT}$ (Figures 5.4F and 5.4G). We failed to observe an increase in favorable contacts between charged interface residues (Figures 5.3B, 5.3D, 5.4D, and 5.4E). Our results suggest that the increased flexibility conferred to $HER3^{E928G}$ at the dimerization interface by adjacent glycine residues (G927 and G928) increases dimerization affinity through backbone H bond optimization.

We next sought to understand the structural basis for potential synergy of $HER3^{E928G}$ with the most common co-occurring HER2 mutants in breast cancer (Figure 5.1D): L755S, V777L, and L869R. Previous studies have shown that HER2 KD mutant monomers, including $HER2^{V777L}$, displayed enhanced kinase activity compared with the $HER2^{WT}$ monomer; HER2 activity was further increased by homodimerization of mutant HER2 compared with the mutant monomer (Bose et al., 2013; Collier et al., 2013). Here, we investigated to what extent these mutations increase stability of the KD active conformation (Figure 5.4A) versus the stability of the asymmetric heterodimer interface (Figure 5.4B). We performed Rosetta DDG calculations of HER2 missense mutations in complex with $HER3^{WT}$ or $HER3^{E928G}$ (Figures 5.4B and 5.4C). The HER2

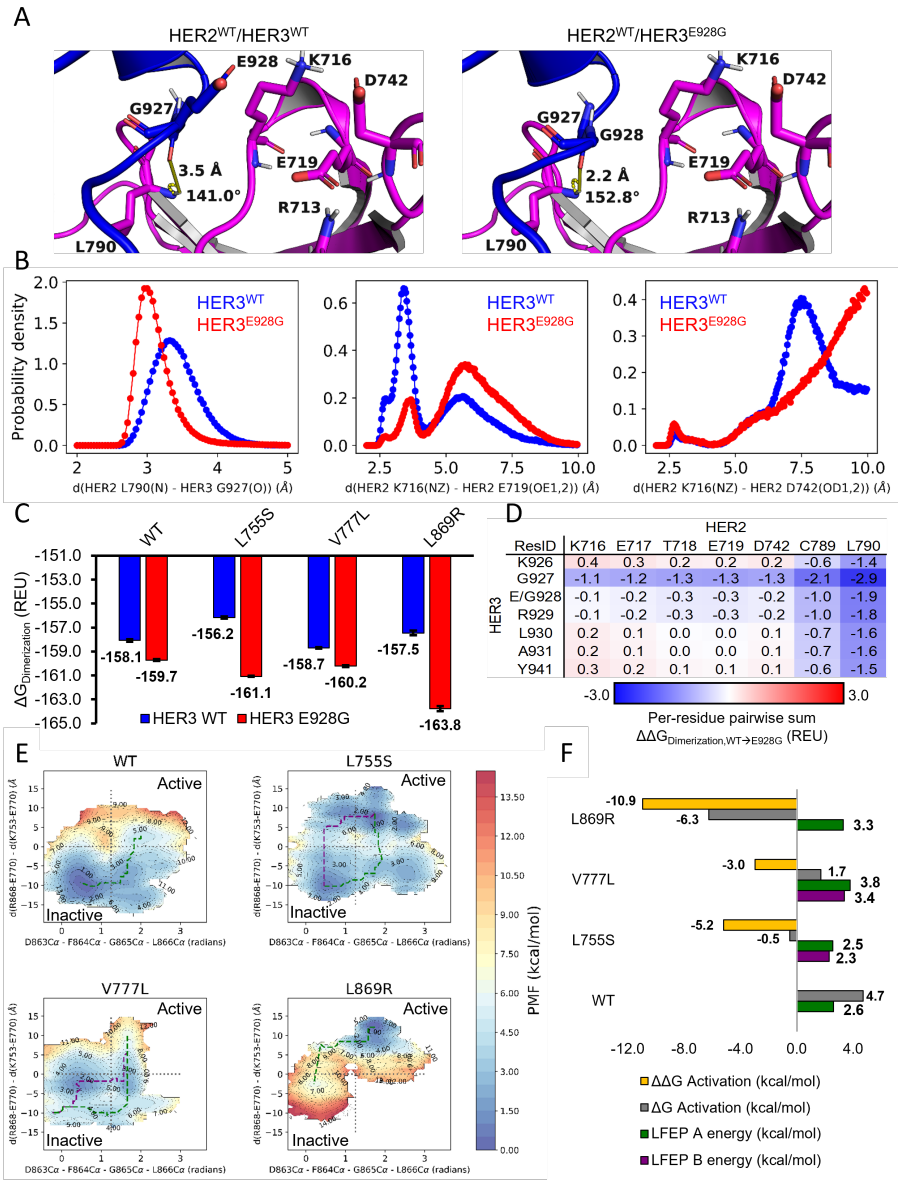


Figure 5.3: Co-occurring HER2/HER3 mutants enhance HER2/HER3 kinase domain association and HER2 kinase activity. (A) Comparison of the computational structural models of the $HER2^{WT}/HER3^{WT}$ and $HER2^{WT}/HER3^{E928G}$ at the asymmetric dimer interface. HER2 is colored purple and HER3 is colored blue. The hydrogen bond between residues G927-O and L790-NH is represented by a yellow line. The hydrogen bond angle given by the L790-N, L790-H, and G927-O atoms is also depicted with a yellow line. (B) Probability density plots of $HER2^{WT}/HER3^{WT}$ and $HER2^{WT}/HER3^{E928G}$ HER3 G927-O – HER2 L790-N hydrogen bond distance (left), HER2 K716-NZ – HER2 E719-OE1,2 bond distance (middle), and HER2 K716-NZ – HER2 D742-OD1,2 bond distance (right). (C) Rosetta HER2/HER3 heterodimerization binding energy. (D) Pairwise sums of per-residue binding energy decomposition for HER2/HER3 heterodimerization. (E) Activation state conformational free energy landscape of $HER2^{WT}$ (upper left quadrant), $HER2^{L755S}$ (upper right quadrant), $HER2^{V777L}$ (lower left quadrant), and $HER2^{L869R}$ (lower right quadrant). (F) Quantification of free energy difference between active and inactive states for each mutant (gray), relative free energy difference compared to $HER2^{WT}$ (yellow), and integration along the lowest free energy path(s) (green and purple).

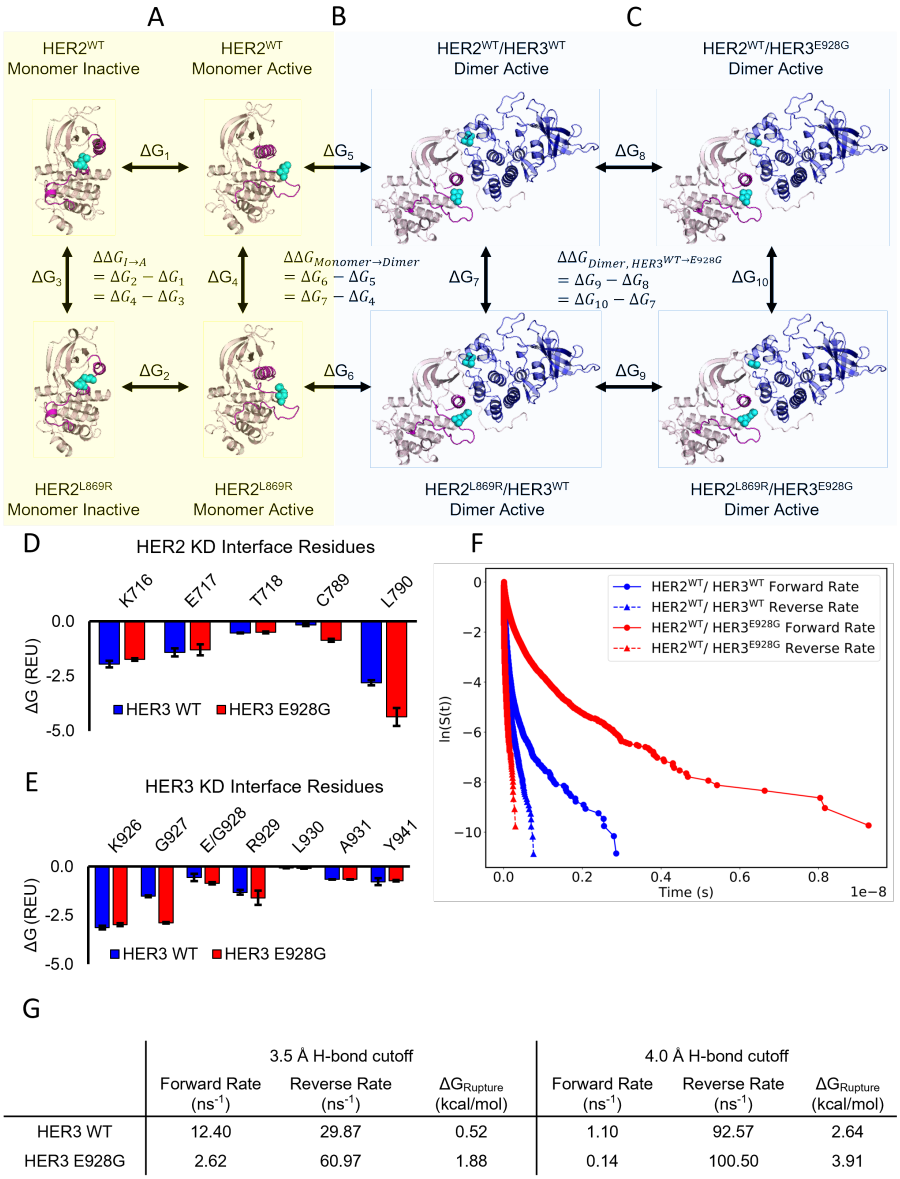
KD mutants did not increase dimerization affinity with $HER3^{WT}$ (Figure 5.3A). In contrast, $HER2^{S310F}$/Y did increase dimerization affinity of the ECDs, potentially because the aromatic side chain of HER2 F/Y310 can make a stable hydrophobic contact with HER3 L272 (Figures 5.5A and 5.5B). $HER3^{E928G}$ enhanced dimerization affinities over $HER3^{WT}$ in all cases (Figures 5.3C and 5.5B).

We tested the hypothesis that HER2 missense mutants increase the stability of the KD active conformation using steered MD and umbrella sampling (US) simulations. We reasoned that mutations that reduce the energetic barrier to activation increase the propensity for dimer formation through conformational selection (Figures 5.4A and 5.4B). $HER2^{WT}$ is more stable in the inactive conformation than the active conformation in our US simulations (Figures 5.3E and 5.3F). In contrast, both $HER2^{L869R}$ and $HER2^{L755S}$ favor the active conformation (Figures 5.3E and 5.3F). Consistent with previous accelerated MD simulations (Robichaux et al., 2019), $HER2^{V777L}$ retained a preference for the inactive conformation in our simulations; however, the barrier to activation is reduced, suggesting that $HER2^{V777L}$ is more readily activated than $HER2^{WT}$. These results suggest that the tested HER2 KD missense mutations lower the free energy barrier between the inactive and active KD conformations, while $HER3^{E928G}$ enhances the stability of the dimerization interface, such that HER2missense/$HER3^{E928G}$ co-mutations cooperatively promote oncogenic activation.

5.2.3 Co-occurring HER2/HER3 mutants enhance ligand-independent HER2/HER3 and PI3K activation

To test our computational predictions, we performed co-immunoprecipitation (co-IP) in HEK293 cells transiently transfected with WT (wild type) or mutant HER2 and HER3. In agreement with the structural predictions (Figures 5.3A and 5.5B), co-expression of $HER3^{E928G}$ enhanced the interaction with $HER2^{S310F}$, L755S, or V777L compared with $HER3^{WT}$ (Figures 5.6A and 5.6B). The stronger association between $HER2^{L755S}$ and $HER3^{E928G}$ compared with either mutant alone was confirmed by proximity ligation assay (PLA) (Figures 5.7A and 5.7B).

Treatment with the HER3 ligand neuregulin (NRG) triggers HER2/HER3 heterodimerization and pathway activation. We asked whether $HER3^{E928G}$ can bypass the effect of NRG stimulation via enhanced interaction with the KD of HER2. Coexpression of $HER3^{E928G}$ with $HER2^{WT}$ strongly enhanced ligand-independent HER3 phosphorylation in serum-starved HEK293 cells (Figure 5.6C) in agreement with previous studies (Jaiswal et al., 2013). Similarly, $HER2^{L755S}$ and $HER2^{V777L}$, when co-expressed with $HER3^{WT}$, increased ligand-independent HER2 and HER3 phosphorylation. Levels of P-HER3 were highest in the double-mutant cells. Similar results were obtained when only the intracellular domains of WT or mutant HER2 and HER3 were expressed (Figure 5.7C). Treatment with NRG was sufficient to stimulate HER2 and HER3 phosphorylation in cells co-expressing $HER2^{WT}$ and $HER3^{WT}$, similar to the effects of HER2/HER3 double mutants



G

	3.5 Å H-bond cutoff			4.0 Å H-bond cutoff		
	Forward Rate (ns^{-1})	Reverse Rate (ns^{-1})	$\Delta G_{Rupture}$ (kcal/mol)	Forward Rate (ns^{-1})	Reverse Rate (ns^{-1})	$\Delta G_{Rupture}$ (kcal/mol)
HER3 WT	12.40	29.87	0.52	1.10	92.57	2.64
HER3 E928G	2.62	60.97	1.88	0.14	100.50	3.91

Figure 5.4: HER2 and HER3 missense mutations enhance receptor heterodimerization with complementary but distinct mechanisms. (A) Thermodynamic cycle relating $HER2^{WT}$ to HER2mutant active to inactive conformational state transition free energy. $HER2^{L869R}$ is displayed as an example of HER2mutant mutants. (B) Thermodynamic cycle relating $HER2^{WT}$ to HER2mutant heterodimerization free energy with $HER3^{WT}$. (C) Thermodynamic cycle relating HER2/$HER3^{WT}$ and HER2/$HER3^{E928G}$ heterodimerization free energies. Here, we evaluated the relative free energies of HER2mutant activation compared to $HER2^{WT}$ (A) with steered MD and umbrella sampling simulations. We evaluated the relative free energies of $HER2^{WT}$ and HER2mutant heterodimerization with $HER3^{WT}$ (B) and $HER3^{E928G}$ (C) with Rosetta. We also utilized conventional MD simulations to investigate differences in heterodimerization affinity of $HER2^{WT}$ with $HER3^{WT}$ vs. $HER3^{E928G}$. (D) Per-residue energy decomposition of select HER2 residues at the HER2/HER3 dimerization interface. (E) Per-residue energy decomposition of select HER3 residues at the HER2/HER3 dimerization interface. All per-residue energies reported as mean +/- standard error across 20 lowest interface energy samples per group. (F) Log-scaled survival curves of the G927 – L790 backbone hydrogen bond rupture event with a 3.5 Å cutoff. (G) Hydrogen bond forward (rupture) and reverse (formation) rates and the free energy associated with hydrogen bond rupture using hydrogen bond distance cutoff values of 3.5 Å or 4.0 Å.

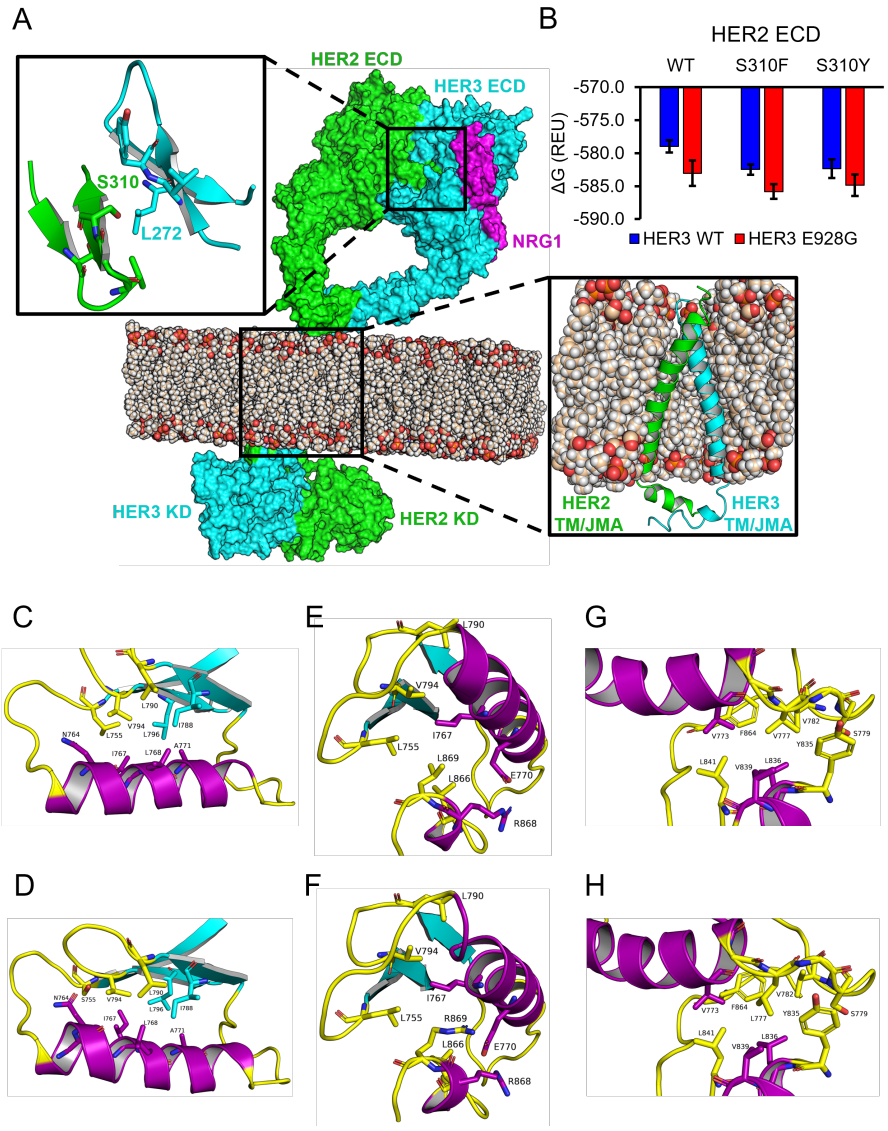


Figure 5.5: Structural features of HER2 missense mutants. (A) Computational structural model of the near full-length HER2^{WT} (green) and HER3^{WT} (cyan) heterodimer with in complex with NRG1 (purple). The modeled heterodimer includes the extracellular domain (ECD; subdomains I – IV), transmembrane domain (TMD), juxtamembrane domain (JMD), and kinase domain (KD) of both HER2 and HER3. The unstructured C-terminal tails were excluded from modeling. (B) Rosetta HER2/HER3 heterodimerization binding energies for the HER2^{S310F} and HER2^{S310Y} mutants with HER3^{WT} and HER3^{E928G}. Reported as mean +/- standard error across 5 lowest interface energy samples per group. (C) HER2^{WT} active state depicting L755 interacting with hydrophobic core residues at the β3-αC interface. (D) HER2^{L755S} active state depicting S755 interacting with hydrophobic core residues at the β3-αC interface. (E) HER2^{WT} inactive state depicting L869 interacting with hydrophobic core. (F) HER2^{L869R} inactive state depicting R869 interacting with hydrophobic core. (G) HER2^{WT} active state depicting V777 interacting with the back hydrophobic pocket. (H) HER2^{V777L} active state depicting L777 interacting with the back hydrophobic pocket.

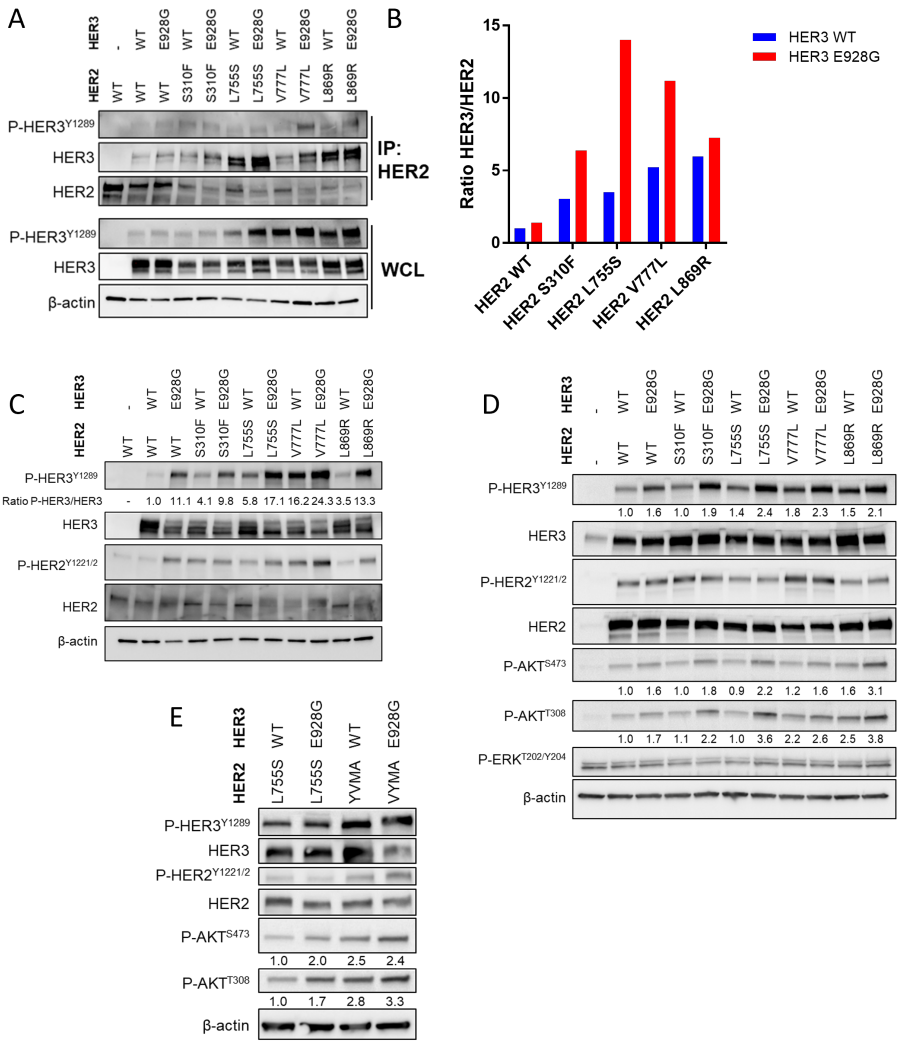


Figure 5.6: HER3^{E928G} enhances HER2/HER3 association and PI3K pathway activation. (A) HEK293 cells were co-transfected with WT or mutant HER2 and HER3^{WT} or HER3^{E928G}. For immunoprecipitation, lysates were incubated with HER2 antibody Ab-17 overnight at 4°C, followed by incubation with Protein G beads and magnetic separation. (B) Immunoblot bands from (A) were quantified using ImageJ. (C) HEK293 cells were co-transfected with WT or mutant HER2 and HER3^{WT} or HER3^{E928G}. Cells were serum-starved overnight, then lysed. Cell lysates were probed with the indicated antibodies. (D) MCF10A cells stably expressing WT or mutant HER2 and HER3^{WT} or HER3^{E928G} were starved in EGF/insulin-free media + 1% CSS overnight. Lysates were probed with the indicated antibodies. (E) MCF10A cells stably expressing the indicated transgenes were starved and lysed as in (D). Where indicated, western blot bands were quantified using ImageJ. The ratios were normalized to the WT/WT condition. Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

in unstimulated cells (Figure 5.6C). These results support a model whereby the concurrent HER2/ HER3 KD mutants promote ligand-independent HER2/HER3 KD association and HER2 kinase activation.

Next, we stably transduced MCF10A breast epithelial cells with WT and mutant HER2, each with WT or mutant HER3. In low-serum conditions, cells expressing the double mutants showed the highest levels of P-HER3 (Figure 5.6D). Unlike HER2, P-HER3 can directly bind to the p85 subunit of PI3K, inducing PI3K activity (Haikala and Janne, 2021). Consistent with this, levels of P-AKT were also highest in double-mutant cells (Figure 5.6D). P-HER3 and P-AKT were enhanced to a similar degree by NRG stimulation in HER2-mutant/HER3^{WT} cells (Figure 5.7D).

The above experiments were performed in the context of ectopic expression of HER2 and HER3; however, most concurrent HER2 and HER3 mutations occur in the absence of HER2 gene amplification (Figures 5.1A and 5.1B). Therefore, we expressed HER3^{WT} or HER3^{E928G} in (1) OVCAR8 ovarian cells, which contain an activating somatic HER2^{G776V} mutation without HER2 amplification (Sudhan et al., 2020), and (2) MCF7 HER2-non-amplified breast cancer cells isogenically modified to express HER2^{L755S} or HER2^{V777L} at endogenous levels (Zabransky et al., 2015). Expression of HER3^{E928G} enhanced co-IP with mutant HER2 in OVCAR8 cells and enhanced P-HER3 in both models compared with HER3^{WT} (Figures 5.6E and 5.7E). Levels of P-AKT were also increased in OVCAR8 cells expressing HER3^{E928G}, but not in MCF7 double-mutant cells, perhaps because these cells harbor an activating PIK3CA mutation. These results suggest that concurrent HER2/HER3 mutants enhance ligand-independent PI3K activity, providing a plausible explanation for the mutual exclusivity of co-occurring HER3 and PIK3CA mutations in HER2-mutant breast cancers (Figure 5.1A).

We noted above that HER2 insertion mutations did not cooccur with HER3 mutations (Figures 5.1A and 5.1B). Therefore, we asked whether the HER2^{Y772_A775dup} (HER2^{YVMA}) insertion mutant could activate HER2/PI3K to a similar degree as cooccurring HER2 and HER3 missense mutants. We modeled the insertion mutants HER2^{YVMA} and HER2^{G778_P780dup} (HER2^{GSP}) mutations based on the HER2^{WT} and EGFR^{D770_N771insNPG} structures (Figure 5.7F). Simulations suggest that HER2^{GSP} and HER2^{YVMA} have reduced free energy barriers to activation relative to HER2^{WT} (Figures 5.7F and 5.7G). Next, we stably transduced MCF10A cells with HER2^{YVMA} and HER3^{WT} or HER3^{E928G}. Both HER2/HER3 co-IP and P-AKT levels were similar in cells expressing HER2^{YVMA}/HER3^{WT} and HER2^{L755S}/HER3^{E928G} (Figures 5.6 and 5.7H). Co-expression of HER3^{E928G} with HER2^{YVMA} did not further increase P-AKT, suggesting that HER2 insertion mutations and HER3 mutations are stronger activators of PI3K than HER2 missense mutations alone.

While HER3^{E928G} is the most common HER3 mutation in breast cancer, we noted several cases of co-occurring HER2/ HER3 ECD mutations (Figure 5.1D). Thus, we expressed each HER3 ECD mutation to-

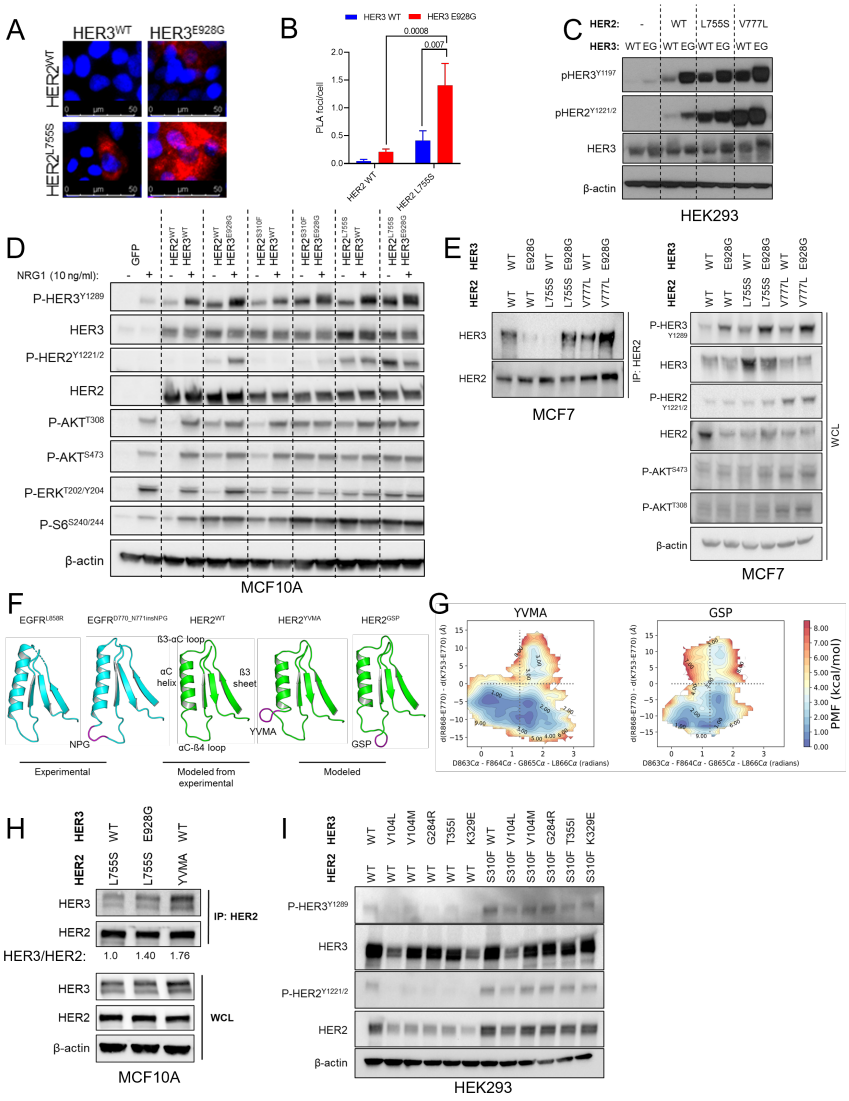


Figure 5.7: Effects of co-occurring HER2/HER3 mutations or HER2 insertion mutations on HER2 kinase activity and HER2/HER3 KD interaction. (A) The intracellular domains (ICDs) of WT or mutant HER2 and HER3 were transiently transfected into HEK-293 cells. Cell lysates were probed with the indicated antibodies. EG, E928G. (F) Illustration of exon 20 insertion mutants. Exon 20 insertion mutations are highlighted in purple. (G) Activation state conformational free energy landscapes of the $HER2^{YVMA}$ and $HER2^{GSP}$ insertion mutants. (D) MCF10A cells stably expressing the indicated genes were cultured in EGF/insulin-free media. Lysates were subjected to immunoprecipitation with the HER2 Ab-17 antibody. Western blot bands were quantified using ImageJ and normalized to the $HER2^{L755S}$/$HER3^{WT}$ condition. (E) HEK293 cells were co-transfected with full-length $HER2^{WT}$ or $HER2^{S310F}$ along with WT or mutant HER3 (ECD mutations). Cells were serum-starved overnight. Cell lysates were probed with the indicated antibodies. Data and illustrations for figure panels A, B, C D, E, H, and I produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

gether with $HER2^{WT}$ or $HER2^{S310F}$ in HEK293 cells. $HER2^{S310F}$ expression with $HER3^{WT}$ resulted in increased ligand-independent HER2 and HER3 phosphorylation compared with $HER2^{WT}$ (Figure 5.7I). However, co-expression of HER3 ECD mutants did not further enhance phospho-HER2 or -HER3, suggesting that these HER3 mutants do not promote ligand-independent HER2/HER3 activation.

5.2.4 Co-occurring HER2/HER3 mutants enhance oncogenic growth and invasion

Next, we asked whether concurrent HER2/HER3 mutants cooperate to transform breast cancer cells. While most of the cooccurring mutations enhanced growth in 2D and 3D (Figures 5.8A and 5.8B), expression of the most common pair, $HER2^{L755S}$/$HER3^{E928G}$, did not further enhance monolayer 2D growth above that of $HER2^{L755S}$ alone.

However, when cultured in 3D Matrigel, MCF10A $HER2^{L755S}$/$HER3^{E928G}$ cells formed large invasive acini in the absence of added NRG1 (Figures 5.8C and 5.8D), suggestive of a more transformed phenotype. Similar invasive acini were formed by cells expressing $HER2^{S310F}$/$HER3^{E928G}$ and $HER2^{L869R}$/$HER3^{E928G}$, but not by cells expressing either HER2 variant with $HER3^{WT}$ (Figure 5.9A). Notably, NRG1 treatment phenocopied the effect of $HER3^{E928G}$ in cells expressing $HER3^{WT}$ and HER2 mutants (Figure 5.8C). Ligand-independent invasive acini were formed by cells transduced with $HER2^{YVMA}$, but this effect was not enhanced by co-transduction with mutant HER3. Invasion through Matrigel-coated chambers was strongly enhanced by all of the double mutants or by $HER2^{YVMA}$/$HER3^{WT}$ (Figures 5.8E, 5.8F, and Figure 5.9B–Figure 5.9E). Together, these results suggest that concurrent HER2/HER3 mutants enhance ligand-independent PI3K pathway activation, which is associated with increased invasion (Samuels et al., 2005).

5.2.5 $HER3^{E928G}$ promotes resistance to HER2-targeting antibodies

We next asked whether HER2- and HER3-targeting antibodies could disrupt the association of $HER3^{E928G}$ with HER2 and the enhanced oncogenicity conferred by co-occurring HER2/HER3 mutations. We used the HER2 antibodies trastuzumab and pertuzumab, which disrupt ligand-dependent and -independent HER2/HER3 dimers (Agus et al., 2002; Junttila et al., 2009) and PanHER, a mixture of antibodies targeting EGFR, HER2, and HER3 that induces ERBB receptor downregulation (Jacobsen et al., 2015). In agreement with previous studies (Greulich et al., 2012; Kavuri et al., 2015), MCF10A cells expressing the extracellular $HER2^{S310F}$ mutation were exquisitely sensitive to the combination of trastuzumab and pertuzumab and to PanHER (Figures 5.10A–5.10C and 5.11A). However, co-expression of $HER3^{E928G}$ reversed this response (Figures 5.10B and 5.10C).

co-IP of cell lysates with HER2 antibodies showed that $HER2^{S310F}$/HERWT dimerization was disrupted by trastuzumab and pertuzumab. In cells expressing $HER2^{S310F}$/$HER3^{E928G}$, dimerization was not affected

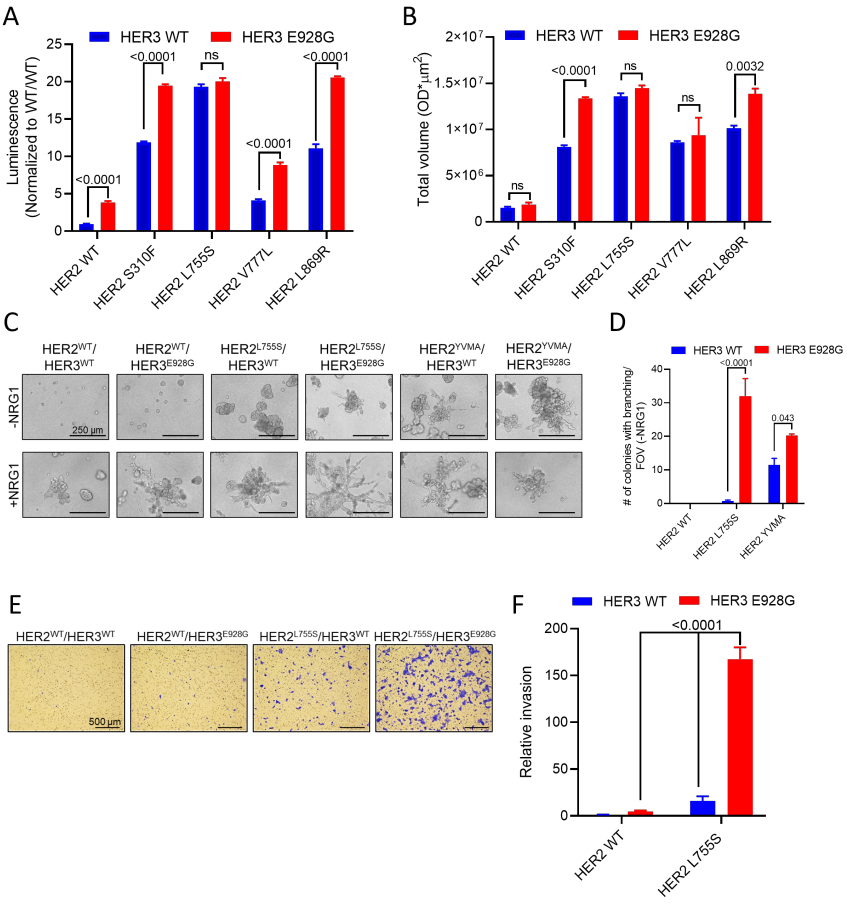


Figure 5.8: Co-occurring HER2/HER3 mutations enhance oncogenic growth and invasion of breast epithelial cells. (A) MCF10A cells stably expressing WT or mutant HER2 and HER3 were grown in 2D in EGF/insulin-free media + 1% CSS for 6 days. Cell viability was measured by Cell Titer Glo. (B) MCF10A cells were grown in 3D Matrigel in EGF-insulin-free media + 1% CSS and stained with MTT. The total volume of colonies per well was quantified using the Gelcount instrument. Data represent the average +/- SEM of three replicates (****, p¡0.0001, one-way ANOVA + Bonferroni multiple comparisons test). (C) MCF10A cells stably expressing WT or mutant HER2 and HER3 were grown in 3D Matrigel in EGF-free media + 1% CSS +/- 10 ng/ml NRG1. (D) The number of colonies showing invasive branching per field of view (FOV) was quantified. Data represent the average +/- SD of three replicates (**, p¡0.01, student t-test). (E) MCF10A cells stably expressing the indicated genes were seeded on Matrigel-coated chambers. After 22 h, invading cells were stained with crystal violet. (F) Relative invasion (normalized to $HER2^{WT}/HER3^{WT}$) from two FOVs per well was quantified using ImageJ. Data represent the average +/- SD of 3-4 replicates (****, p¡0.0001, One-way ANOVA + Bonferroni multiple comparisons test). Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

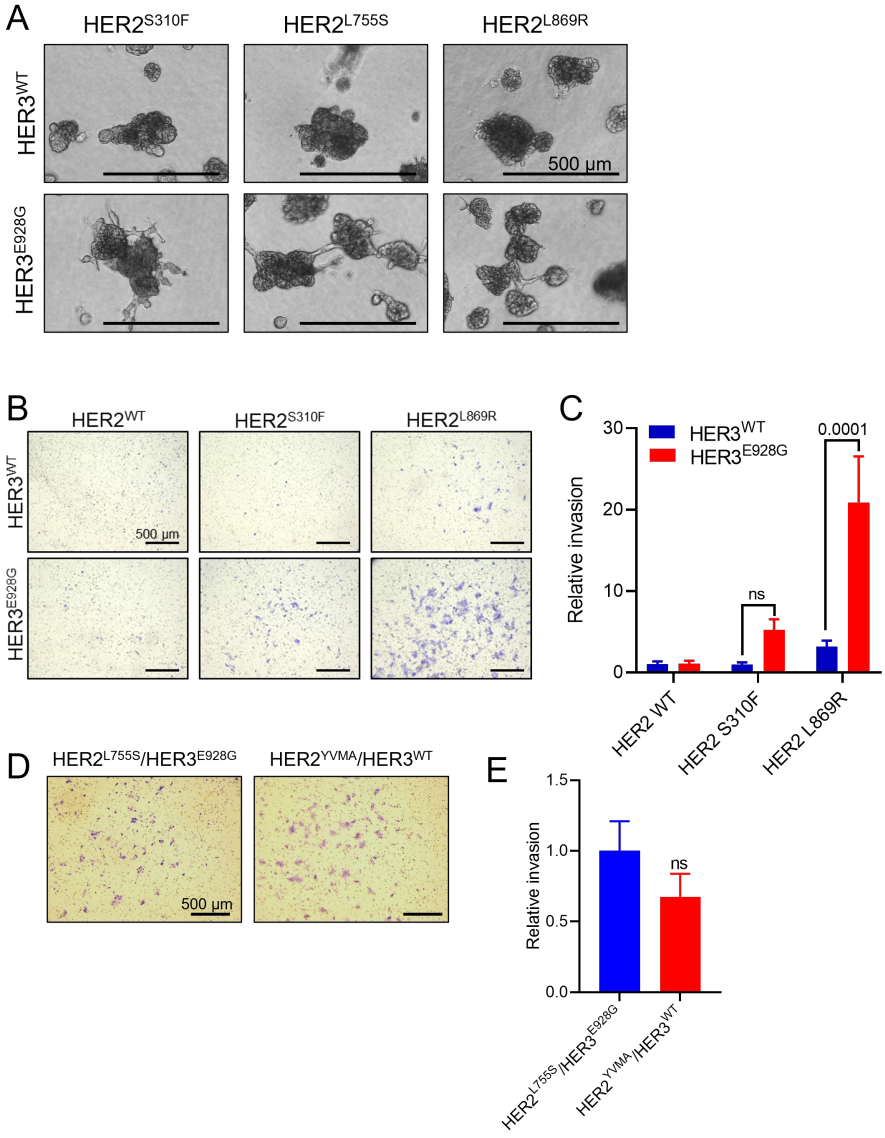


Figure 5.9: Co-occurring HER2/HER3 missense mutations or HER2 insertion mutations increase the invasive capacity of breast epithelial cells. (A) MCF10A cells stably expressing the indicated genes were grown in 3D Matrigel in EGF-free media + 1% CSS. (B) MCF10A cells stably expressing the indicated genes were seeded on Matrigel-coated chambers. After 22 h, invading cells were stained with crystal violet. (C) Relative invasion (normalized to $HER2^{WT}/HER3^{WT}$) from two FOVs per well was quantified using ImageJ. Data represent the average +/- SEM (n¿3). P values, two-way ANOVA + Bonferroni. (D) MCF10A cells stably expressing the indicated genes were seeded on Matrigel-coated chambers and stained as in (B). (E) Relative invasion (normalized to $HER2^{L755S}/HER3^{E928G}$ was quantified as in (C). Data represent the average +/- SEM (n¿4). Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

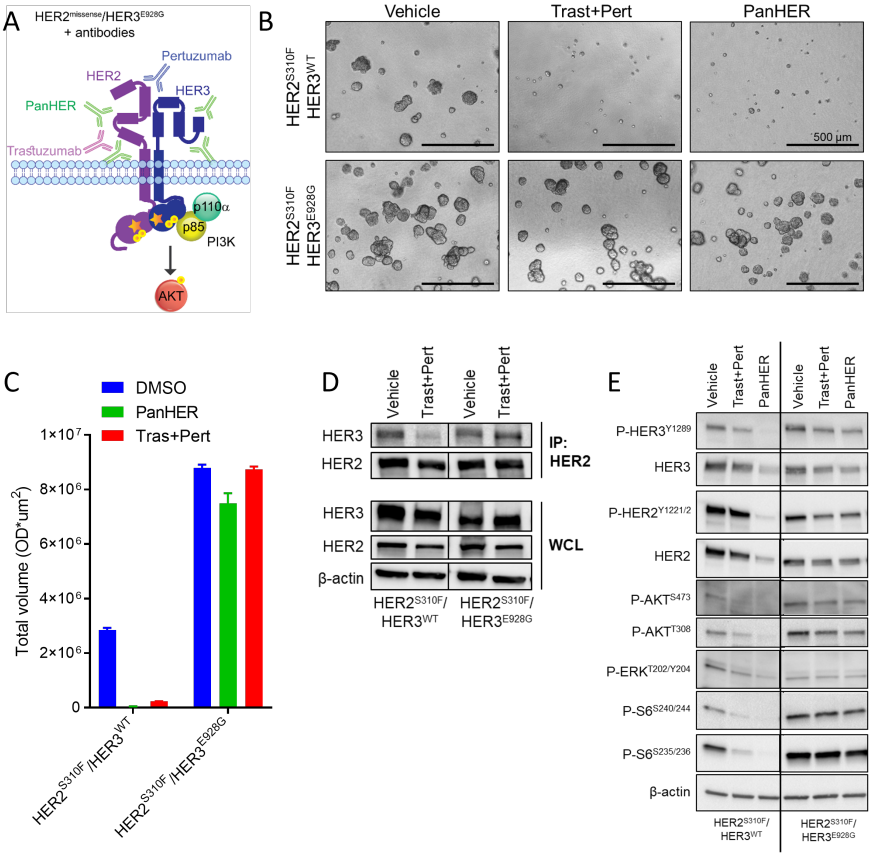


Figure 5.10: HER3^{E928G} promotes resistance to HER2- and HER3-targeting antibodies by retaining HER2/HER3 kinase domain association. A) Model of HER2/HER3^{E928G} heterodimer bound to trastuzumab, pertuzumab, PanHER antibody mixture, or LJM716. The enhanced kinase domain association mediated by HER3^{E928G} is not predicted to be disrupted by antibodies blocking the associationof the HER2 and HER3 ECDs. (B) MCF10A cells stably expressing the indicated genes were grown in 3D Matrigel in EGF/insulin-free media treated with vehicle (PBS), 20 g/ml PanHER, 20 g/ml each trastuzumab + pertuzumab and stained with MTT. (C) The total volume of colonies per well was quantified using the Gelcount instrument. Data represent the average +/- SD of three replicates. (D) MCF10A cells stably expressing HER2^{S310F}/HER3^{WT} or HER2^{S310F}/HER3^{E928G} were treated with vehicle (PBS) or 20 g/ml each trastuzumab and pertuzumab for 24 h in EGF/insulin-free media + 1% CSS. Following an acid wash to remove bound antibodies, HER2 immunoprecipitation was performed as described in STAR Methods. (E) MCF10A cells stably expressing HER2^{S310F}/HER3^{WT} or HER2^{S310F}/HER3^{E928G} were treated with vehicle (PBS), 20 g/ml each trastuzumab and pertuzumab, or 20 g/ml PanHER for 24h in EGF/insulin-free media + 1% CSS. Lysates were probed with the indicated antibodies. Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

by antibody treatment (Figure 5.10D). Similarly, the antibodies blocked P-HER3, P-AKT, and the downstream effector P-S6 in MCF10A cells expressing $HER2^{S310F}/HER3^{WT}$, but failed to do so in cells expressing $HER2^{S310F}/HER3^{E928G}$ (Figure 5.10E). Flow cytometry analysis revealed that $HER3^{E928G}$ did not disrupt trastuzumab binding to cell surface HER2 (Figure 5.11B). These results suggest that $HER3^{E928G}$ may enable the intracellular association of HER2 and HER3 KD mutants, even when the ECD interaction is disrupted by neutralizing antibodies.

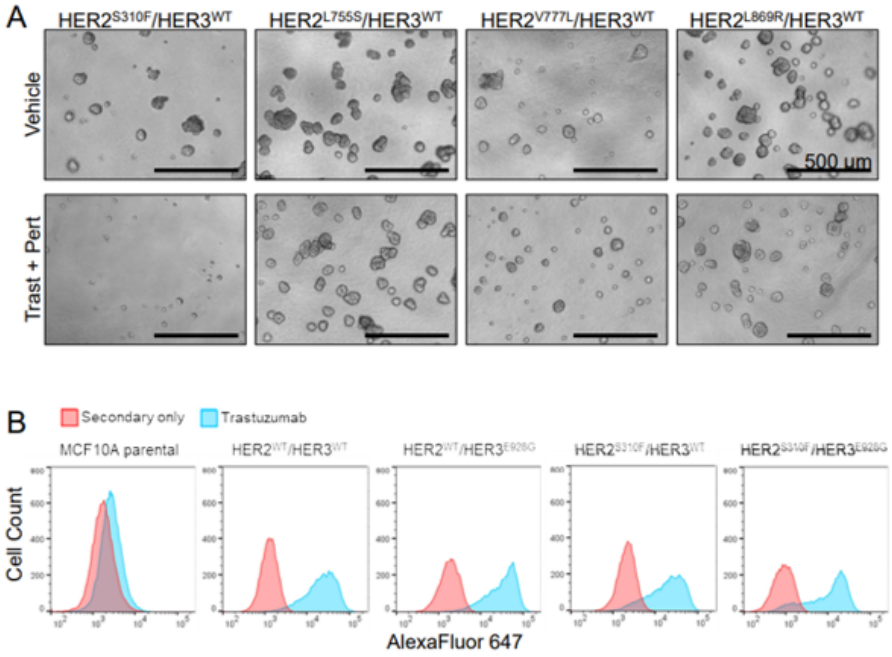


Figure 5.11: $HER2^{S310F}$-induced transformation is blocked by anti-HER2 antibodies. (A) MCF10A cells stably expressing the indicated genes were grown in 3D Matrigel in EGF/insulin-free media treated with vehicle (PBS) or 20 g/ml each trastuzumab + pertuzumab for 7 d. Scale bar, 500 m. (B) MCF10A cells stably expressing the indicated transgenes were stained with 0.2 g/ml trastuzumab and an Alexa Fluor 647-conjugated goat anti-human IgG secondary antibody and analyzed by flow cytometry. Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

5.2.6 $HER3^{E928G}$ modulates sensitivity to neratinib

The HER2 TKI neratinib has emerged as a promising treatment for HER2-mutant metastatic breast cancer. However, only a subset of HER2-mutant patients respond to neratinib (Hyman et al., 2018; Ma et al., 2017; Smyth et al., 2020). Therefore, we asked whether concurrent $HER3^{E928G}$ mutations affect the ability of neratinib to inhibit HER2. Neratinib is an ATP-competitive TKI, so its efficacy is a function of ATP-binding affinity. MD simulations and molecular mechanics generalized Born and surface area binding energy calculations of the $HER2^{WT}$-ATP complex heterodimerized with $HER3^{WT}$ or $HER3^{E928G}$ suggest that $HER3^{E928G}$

enhanced binding affinity to ATP (Figure 5.12A). Similar results were seen in simulations of missense variants (Figures 5.12B and 5.12C). Our simulations suggest that $HER3^{E928G}$ reduces the binding affinity of neratinib to $HER2^{WT}$, $HER2^{L755S}$, and $HER2^{L869R}$ (Figure 5.12D). They also suggest that $HER2^{L755S}$, and to a lesser extent $HER2^{L869R}$, may have reduced sensitivity to neratinib that is compounded by co-occurrence with $HER3^{E928G}$, consistent with previous reports that $HER2^{L755S}$ may be less sensitive to HER2 TKIs (Li et al., 2019; Robichaux et al., 2019). In contrast, $HER2^{V777L}$ is expected to mostly retain sensitivity to neratinib even when co-occurring with $HER3^{E928G}$ (Figure 5.12D).

We subsequently tested the neratinib sensitivity of MCF10A cells co-expressing WT or mutant HER2 and HER3. Co-expression of $HER3^{E928G}$ resulted in a 15-fold shift in neratinib halfmaximal inhibitory concentration (IC50) in MCF10A $HER2^{S310F}$expressing cells (Figure 5.12E). Similar results were obtained with other HER2 TKIs (poziotinib, afatinib, and tucatinib), suggesting that expression of $HER3^{E928G}$ reduces sensitivity to most HER2 ATP-competitive inhibitors (Figure 5.13A). However, the shift in IC50 varied in a HER2 allele-specific manner (Figures 5.12F and 5.13B), consistent with our computational predictions (Figures 5.12D). For example, $HER2^{L755S}$ cells were less sensitive to neratinib compared with $HER2^{S310F}$, consistent with previous reports (Li et al., 2019; Robichaux et al., 2019). This trend was similar in 3D Matrigel cultures: treatment with neratinib blocked growth of MCF10A $HER2^{S310F}$/$HER3^{WT}$ and $HER2^{V777L}$/$HER3^{WT}$ cells and partially blocked growth of MCF10A $HER2^{L869R}$/$HER3^{WT}$ cells, whereas cells expressing $HER2^{L755S}$ were largely resistant (Figure 5.12G). Co-expression of $HER3^{E928G}$ reduced the response to neratinib in cells expressing most HER2 mutants. Consistent with the effects on cell growth, neratinib treatment blocked P-HER3, P-AKT, and P-S6 in MCF10A cells expressing HER2mutant/$HER3^{WT}$, but to a lesser degree in cells expressing $HER2^{L755S}$/$HER3^{WT}$, while neratinib failed to block HER3/PI3K signaling in cells expressing $HER3^{E928G}$ (Figure 5.13C). Furthermore, OVCAR8 cells (somatic $HER2^{G776V}$) ectopically expressing $HER3^{E928G}$ (Figure 5.6E) exhibited reduced sensitivity to neratinib compared with cells expressing $HER3^{WT}$ (Figure 5.13D).

Next, we established organoids from an HER2-mutant, nonamplified breast tumor model: the SA493 patient-derived xenograft (PDX), derived from an ER+/$HER2^{S310F}$ lobular breast cancer (Eirew et al., 2015). We confirmed that the organoids retained the $HER2^{S310F}$ mutation (Figure 5.13E). Next, we stably transduced these organoids with $HER3^{WT}$ or $HER3^{E928G}$ (Figure 5.13F); expression of $HER3^{E928G}$ in these HER2-mutant organoids increased P-HER3, P-AKT, and P-S6 (Figure 5.13G). In ligand-free media, cells expressing $HER3^{E928G}$ formed larger, less-organized organoids compared with those expressing $HER3^{WT}$, suggesting that $HER3^{E928G}$ promotes a more aggressive phenotype of this HER2-mutant breast cancer model (Figure 5.13H). While parental organoids and those expressing $HER3^{WT}$ were quite sensitive to trastuzumab + pertuzumab, neratinib, or the combination, organoids expressing $HER3^{E928G}$ exhibited markedly reduced

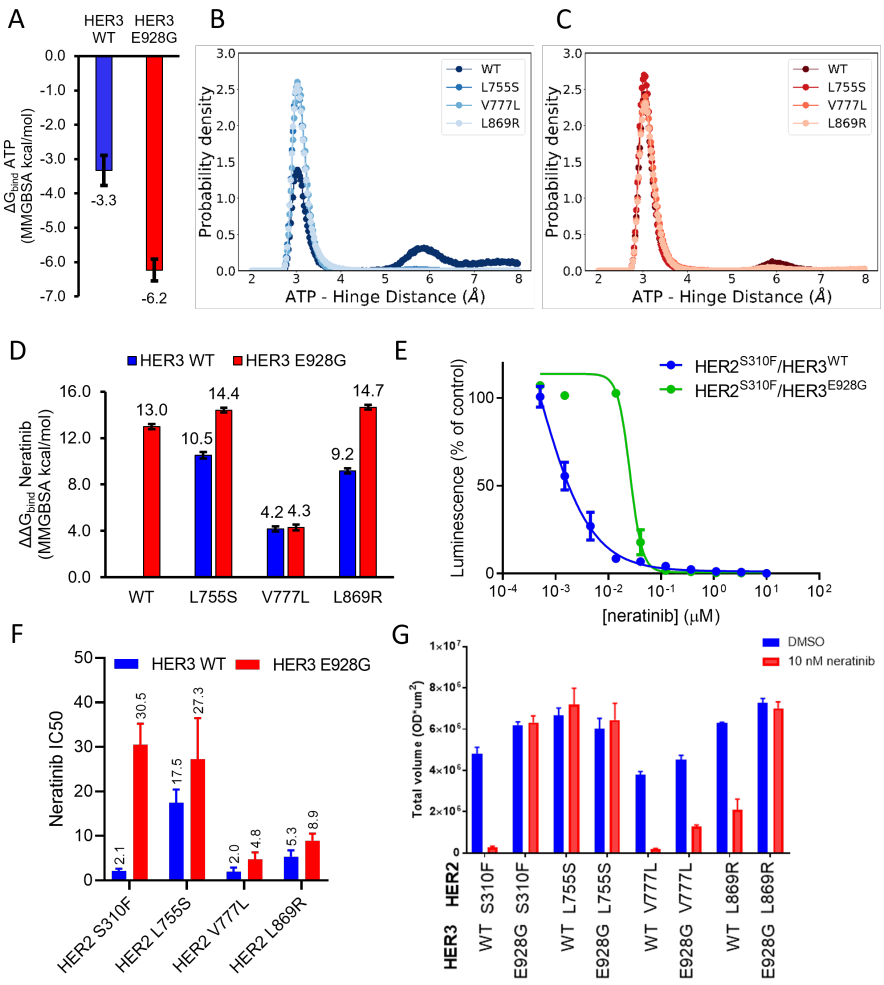


Figure 5.12: Co-occurring HER3 mutations modulate neratinib sensitivity in HER2-mutant cells. (A) Molecular dynamics MM/GBSA binding affinity estimates of ATP to $HER2^{WT}/HER3^{WT}$ and $HER2^{WT}/HER3^{E928G}$. (B) Probability density kinase domain hinge – ATP hydrogen bond distance in $HER2^{WT}$, $HER2^{L755S}$, $HER2^{V777L}$, and $HER2^{L869R}$ dimerized with $HER3^{WT}$. (C) Probability density kinase domain hinge – ATP hydrogen bond distance in $HER2^{WT}$, $HER2^{L755S}$, $HER2^{V777L}$, and $HER2^{L869R}$ dimerized with $HER3^{E928G}$. (D) Molecular dynamics MM/GBSA relative binding affinity estimates of neratinib to different HER2 missense mutants heterodimerized with either $HER3^{WT}$ or $HER3^{E928G}$. (E) MCF10A cells stably expressing the indicated genes were grown in EGF/insulin-free media + 1% CSS and treated with the indicated concentrations of neratinib for 6 days. Cell viability was measured using CellTiterGlo. (F) Neratinib IC50s were determined as in (E). Data represent the average of 3 independent dose-response curves containing 4 replicates each. (G) MCF10A cells stably expressing WT or mutant HER2 and HER3 were grown in 3D Matrigel in EGF-free media + 1% CSS ± 10 nM neratinib and stained with MTT. The total volume of colonies per well was quantified using the Gelcount instrument. Data represent the average ± SD of three replicates. Data and illustrations for figure panels D - G produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

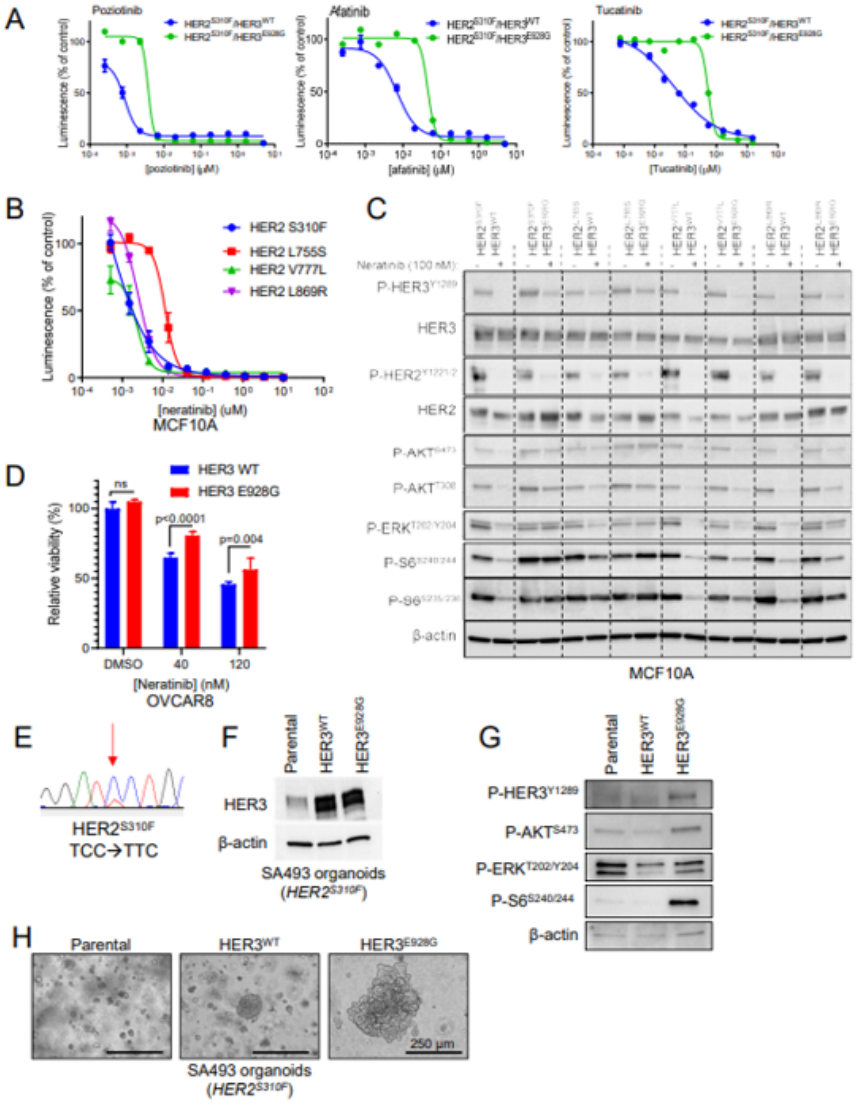


Figure 5.13: The growth of CW2 HER2^{L755S}/HER3^{E928G} colon cancer cells depends on HER2^{L755S} and HER3. A) Electropherograms of ERBB2 cDNA from CW2 cells, indicating heterozygous expression of HER2^{L755S} and HER3^{E928G}. A reverse primer was used for HER2 sequencing. (B) CW2 cells were transfected with siControl or siRNA specifically targeting HER2^{L755S}. qRT-PCR was performed using primers specific for HER2^{WT} (black) or HER2^{L755S} (blue). **, p¡0.01, two-way ANOVA + Bonferroni multiple comparisons test. (C) CW2 cells were transfected control or HER3 siRNA. qRT-PCR was performed using HER3 primers. (D) CW2 cells were transfected with the indicated siRNA and lysed after 48h. Lysates were probed with the indicated antibodies. (E) CW2 cells were transfected with the indicated siRNA. Cell viability after 4 days was measured using the CyQuant assay. **, p¡0.01; ***, p¡0.001, one-way ANOVA + Bonferroni. (F) CW2 cells were transfected with the indicated siRNA. Total cell number was measured after 4 days using a Coulter counter. ***, p¡0.001; ****, p¡0.0001, one-way ANOVA + Bonferroni. Data represent the average ± SD of three independent experiments. Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

sensitivity to these agents (Figure 5.12H). Together, our results suggest that HER3^{E928G} increases ligand-independent growth and reduces sensitivity to HER2-targeting agents in multiple HER2-mutant tumor models.

5.2.7 Cancer cells with co-occurring HER2/HER3 mutations are sensitive to combined inhibition of HER2 and PI3Kα

Our results suggest that HER2/HER3 co-mutations hyperactivate the PI3K/AKT pathway and result in relative resistance to HER2-targeted therapies. Therefore, we tested the combination of neratinib with a PI3K inhibitor in MCF10A cells expressing the double mutants. The combination of neratinib with the PI3Ka inhibitor alpelisib or with the pan-PI3K inhibitor buparlisib blocked P-AKT and P-S6 in MCF10A HER2^{L755S}/HER3^{E928G} and HER2- YVMA cells more potently than either drug alone (Figure 5.14A).

The combination of neratinib and alpelisib also strongly reduced colony growth and invasive acini formation in 3D Matrigel by these cells (Figures 5.14B and 5.14C). Next, we examined CW2 colorectal cancer cells, which harbor somatic HER2^{L755S}/HER3^{E928G} mutations (Figure 5.15A) (Kloth et al., 2016). Small interfering RNA (siRNA)-induced knockdown of either HER2^{L755S} or HER3 showed that the proliferation and PI3K activity in these cells is partially dependent on both mutant HER2 and HER3 (Figures 5.15B–5.15F). The combination of neratinib and alpelisib was required to eliminate P-AKT and synergistically blocked proliferation in these cells (combination index = 0.42) (Figures 5.14D and 5.14E). While 4 h treatment with neratinib + alpelisib strongly blocked P-ERK and P-S6 in CW2 and MCF10A HER2^{L755S}/HER3^{E928G} cells, a rebound was seen at 24 h of treatment (Figures 5.15G and 5.15H), perhaps reflecting activation of feedback pathways (Chakrabarty et al., 2012; Chandarlapaty et al., 2011).

In addition, the combination delayed growth of CW2 xenografts more potently than each drug alone (Figures 5.14F and 5.15I). Together, our data suggest that addition of a PI3Ka inhibitor increases the sensitivity of tumors with HER2mut/HER3^{E928G} to HER2 TKIs.

5.3 Discussion

Somatic HER2 mutations are increasingly being recognized as targetable alterations in breast and other cancers (Mishra et al., 2017; Cocco et al., 2018), prompting a number of studies testing HER2 TKIs in HER2-mutant cancers (Hyman et al., 2018; Robichaux et al., 2019; Smyth et al., 2020). Here, we investigated the intriguing co-occurrence of mutations in HER2 and HER3, genes that encode members of the same signaling complex. We reasoned that such patterns of co-occurrence indicate a selective advantage conferred by both oncogenes during tumor evolution. Recent studies have found that a number of oncogenes, including HER2, HER3, and PIK3CA, often harbor more than one mutation in the driver oncogene, termed 'composite muta-

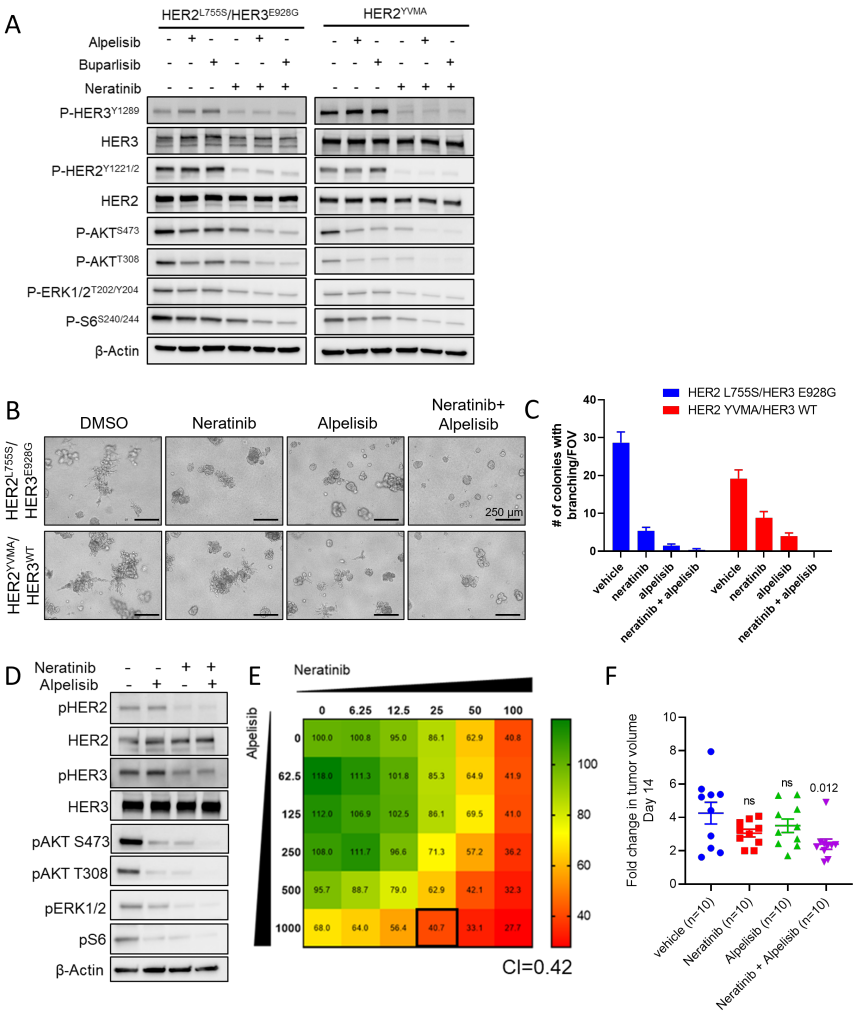


Figure 5.14: Cancer cells harboring co-occurring mutations in HER2 and HER3 are sensitive to combined inhibition of HER2 and PI3Kα. (A) MCF10A cells stably expressing HER2^{L755S}/HER3^{E928G} or HER2^{YVMA}/HER3^{WT} were treated with vehicle (DMSO), 500 nM neratinib, 500 nM buparlisib, 50 nM neratinib, or the indicated combinations for 4 h in EGF/insulin-free media + 1% CSS. Lysates were probed with the indicated antibodies. (B) MCF10A cells stably expressing the indicated genes were grown in 3D Matrigel in EGF/insulin-free media + 1% CSS treated with vehicle (DMSO), 20 nM neratinib, 1 M alpelisib, or the combination. (C) The number of colonies showing invasive branching per field of view (FOV) from (B) was quantified. Data represent the average ± SD of three replicates. (D) CW2 colon cancer cells (HER2^{L755S}/HER3^{E928G}) were treated with vehicle (DMSO), 500 nM alpelisib, 50 nM neratinib, or the combination in serum-free media for 4 h. Lysates were probed with the indicated antibodies. (E) CW2 cells were treated with increasing concentrations of neratinib (0-100 nM) or alpelisib (0-1000 nM) alone or in combination for 72 h. Cell viability was quantified using the CyQuant assay and combination indices were determined using the Chou-Talalay test. Numbers inside each box represent the average % viability (relative to untreated controls) from two independent experiments. (F) Mice carrying CW2 xenografts were treated with vehicle, 40 mg/kg neratinib, 40 mg/kg alpelisib, or the combination for 14 days, starting when tumors reached 200 mm3. Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

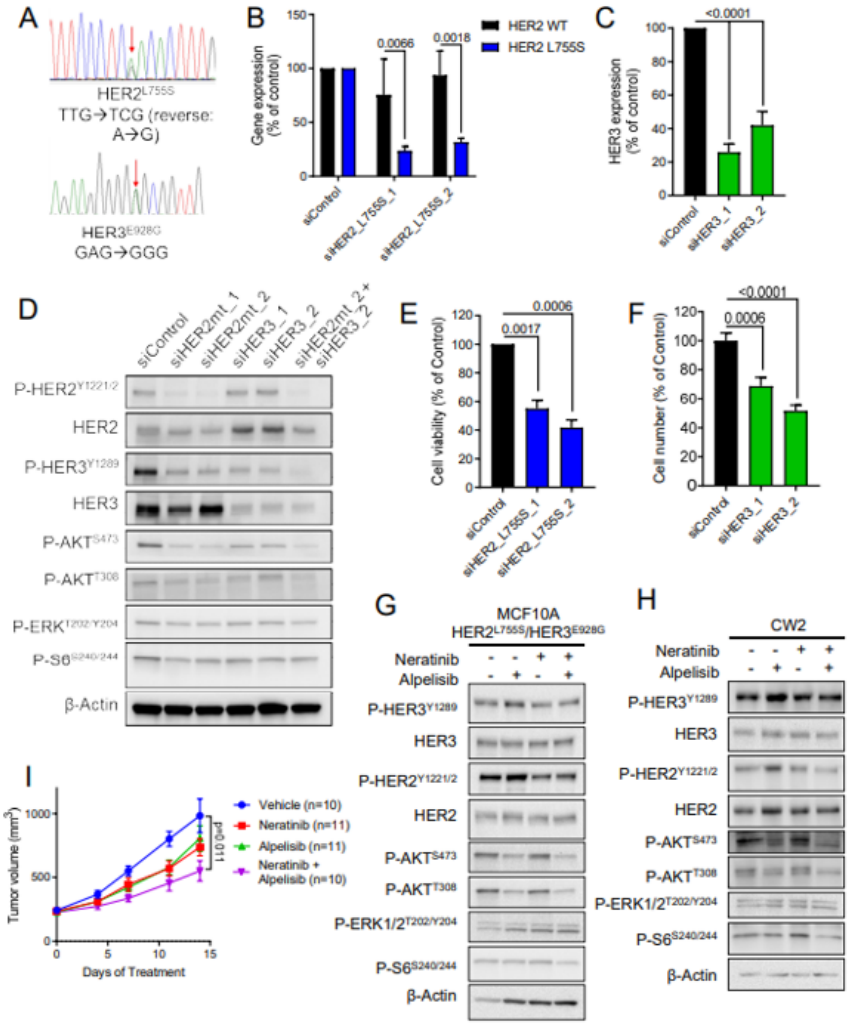


Figure 5.15: The growth of CW2 HER2^{L755S}/HER3^{E928G} colon cancer cells depends on HER2^{L755S} and HER3. (A) Electropherograms of ERBB2 cDNA from CW2 cells, indicating heterozygous expression of HER2^{L755S} and HER3^{E928G}. A reverse primer was used for HER2 sequencing. (B) CW2 cells were transfected with siControl or siRNA specifically targeting HER2^{L755S}. qRT-PCR was performed using primers specific for HER2^{WT} (black) or HER2^{L755S} (blue). P values, two-way ANOVA + Bonferroni. (C) CW2 cells were transfected control or HER3 siRNA. qRT-PCR was performed using HER3 primers. P values, one-way ANOVA + Bonferroni. (D) CW2 cells were transfected with the indicated siRNA and lysed after 48h. Lysates were probed with the indicated antibodies. (E) CW2 cells were transfected with the indicated siRNA. Cell viability after 4 d was measured using the CyQuant assay. P values, one-way ANOVA + Bonferroni. Data represent the average ± SD of three independent experiments. (F) CW2 cells were transfected with the indicated siRNA. Total cell number was measured after 4 d using a Coulter counter. P values, one-way ANOVA + Bonferroni. Data represent the average ± SD of three independent experiments. (G,H) MCF10A HER2^{L755S}/HER3^{E928G} (G) and CW2 (H) cells were treated with vehicle (DMSO), 500 nM alpelisib, 50 nM neratinib, or the combination in serum-free media for 24 h. Lysates were probed with the indicated antibodies. (I) Mice carrying CW2 xenografts were treated with vehicle, 40 mg/kg neratinib, 30 mg/kg alpelisib, or both drugs for 14 d, starting when tumors reached 200 mm3. Data represent the average tumor volme ± SEM. P value, student's t-test, vehicle vs. combination (Day14). Data and illustrations produced by Hanker, A. B., Marín, A.; Jayanthan, H. S.; Ye, D.; Lin, C.-C.; Akamatsu, H.; Lee, K.-M.; Chatterjee, S.; Sudhan, D. R.; Servetto, A.; Brewer, M. R.; Koch, J. P.; He, J.; Lalani, A. S.; and Arteaga, C. L.

tions' (Gorelick et al., 2020; Saito et al., 2020). In particular, composite PIK3CA mutations have been shown to increase PI3K activity and PI3K-dependent tumor growth (Vasan et al., 2019). We speculate that single gain-of-function missense mutations may not fully maximize HER2/HER3 activation, such that either composite HER2 mutations, or co-occurring HER2/HER3 mutations, increase pathway activation and provide a selective advantage.

It is well established that HER2-driven transformation, invasion, and metastasis depends on HER3/PI3K signaling (Holbro et al., 2003; Smirnova et al., 2012; Xue et al., 2006). In addition, activating mutations PIK3CA cooperate with amplified WT HER2, enhancing invasion and metastasis (Chakrabarty et al., 2010; Hanker et al., 2013). In line with these data, co-mutant HER2/HER3 hyperactivate PI3K/AKT and enhance transformation/invasion (Figures 5.6 and 5.8), potentially explaining the observed mutual exclusivity of these alterations in HER2-mutant breast tumors (Figure 5.1A). While clinical information of patients with co-occurring HER2/HER3 mutations is scarce, future studies should address whether this genomic subset of patients correlates with increased metastasis.

We observed strong concordance between our computational structural predictions and biological results. Our simulations suggest that co-occurring HER2 and HER3 mutants enhance the coupling of the receptor KDs, such that HER2 missense mutants increase kinase conformational activation relative to $HER2^{WT}$, while $HER3^{E928G}$ enhances heterodimerization affinity (Figure 5.16B). This model is supported by co-IP, PLA, and immunoblot assays (Figures 5.6 and 5.7). Our simulations also predicted that $HER2^{L755S}$ binds neratinib with reduced affinity (Figure 5.12D). Indeed, $HER2^{L755S}$ was less sensitive to neratinib than the other HER2 mutants in our cell viability and 3D Matrigel assays (Figures 5.12F, 5.12G, 5.13B, and 5.13C), consistent with previous reports (Li et al., 2019; Robichaux et al., 2019). Likewise, our computational modeling predicted that neratinib binding depends on the specific HER2 mutation within the HER2/$HER3^{E928G}$ heterodimer (Figure 5.12D). This was confirmed in cell-based assays: while $HER3^{E928G}$ strongly reduced neratinib sensitivity and neratinib binding in the absence of HER2 KD mutations (i.e., $HER2^{S310F}$/$HER3^{E928G}$), the $HER2^{V777L}$/$HER3^{E928G}$ double mutant retained a strong interaction with neratinib and a high degree of sensitivity to neratinib (Figures 5.12F and 5.12G). Thus, $HER3^{E928G}$ reduces sensitivity to neratinib in a HER2 allele-specific manner.

Our results suggest that HER2 allele-specific differences in neratinib sensitivity are related to unique mechanisms of activation of each mutant. We hypothesize that $HER2^{L755S}$ stabilizes the N-terminal region of the αC helix (Figures 5.5C and 5.5D). In contrast, we hypothesize that $HER2^{V777L}$ increases hydrophobic contacts in the back hydrophobic pocket, but may also function similar to KD insertion mutants (Figures 5.5E and 5.5F). Because L755S more rigidly pulls the αC helix inward from the N-terminal region, the force applied perpendicularly to the αC helix by the neratinib pyridine ring may be greater than in V777L,

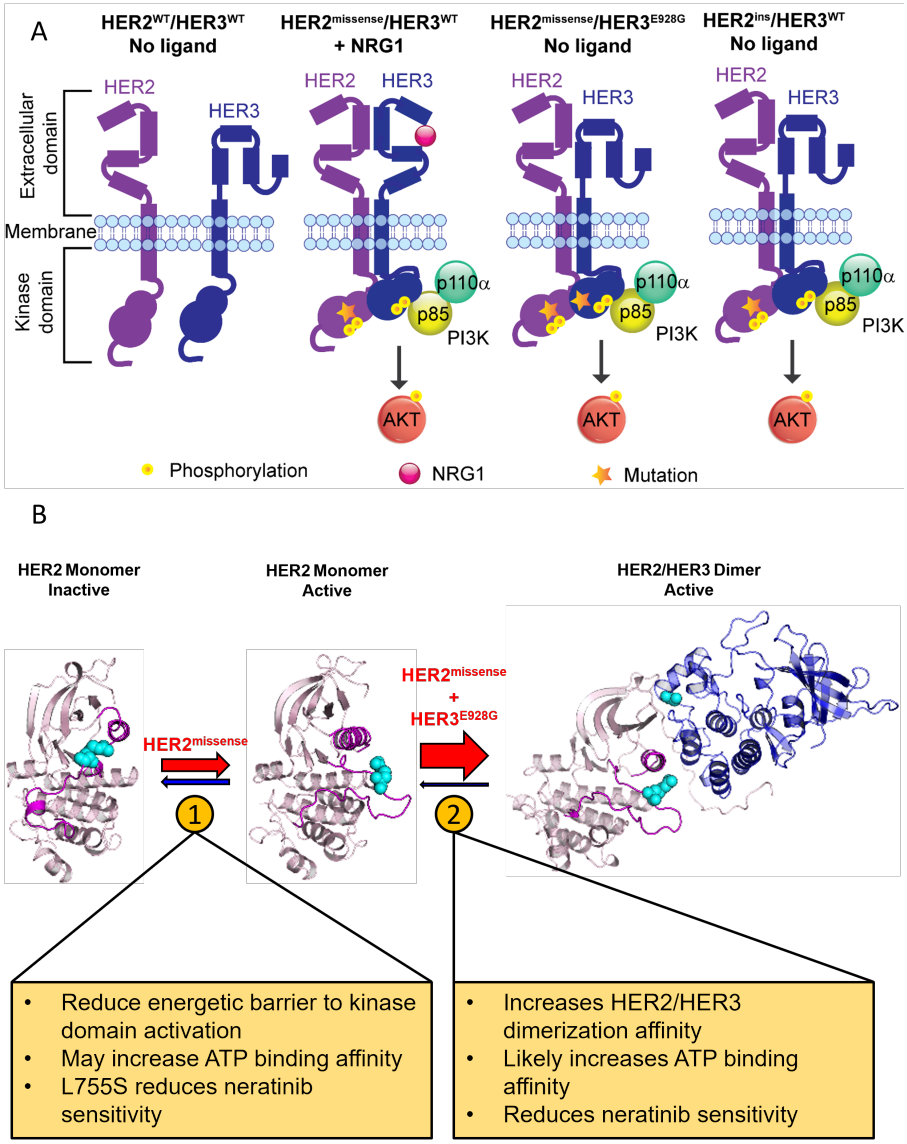


Figure 5.16: Model of HER2/PI3K pathway activation by co-occurring HER2/HER3 mutations. In the absence of ligand, $HER3^{WT}$ is in the closed conformation and does not interact with $HER2^{WT}$. NRG1 treatment (hot pink circle) promotes HER2/HER3 heterodimerization, and a HER2 missense mutation further increases HER3 phosphorylation to recruit the p85 subunit of PI3K and activate PI3K signaling. In the absence of ligand, the $HER3^{E928G}$ mutation phenocopies NRG1 treatment by increasing HER2/HER3 association via enhanced binding of the HER2/HER3 kinase domains, leading to constitutive activation of PI3K. HER2 insertion mutations alone, without HER3 mutations, also increase ligand-independent HER2/HER3 association and PI3K activation. (B) A schematic equilibrium model showing how HER2missense mutations cooperate with $HER3^{E928G}$ to enhance receptor heterodimerization and drive oncogenic activation.

analogous to EGFRL858R (Sogabe et al., 2012). Finally, we hypothesize that HER2^{L869R} decreases the stability of the KD inactive conformation. The intermediate neratinib sensitivity of HER2^{L869R} may be the result of increased occupancy of the active conformation without direct stabilization of the αC helix (Figures 5.5G and 5.5H). Crystallographic studies coupled with detailed structure-activity relationship profiling and long-timescale MD simulations are needed to fully elucidate the structural basis of TKI sensitivity/resistance.

In recent clinical trials of neratinib in patients with HER2- mutant cancer, patients with concurrent HER3 mutations in their tumors exhibited a lower clinical response and shorter progression-free survival (Hyman et al., 2018; Smyth et al., 2020). Our results provide evidence that HER3^{E928G} confers reduced sensitivity to neratinib in HER2-mutant breast cancer cells. In addition to reducing neratinib sensitivity, we found that expression of HER3^{E928G} strongly promoted resistance to HER2- and HER3-targeting antibodies (trastuzumab + pertuzumab or PanHER; Figure 5.12B). Similarly, (Jaiswal et al., 2013) found that HER3^{E928G} was insensitive to HER2- and HER3-targeting antibodies. We predict that small molecules that block HER2/ HER3 KD association would be most likely to block the oncogenic effects of concurrent HER2^{missense}/HER3^{E928G} mutations. To the best of our knowledge, clinical compounds that disrupt HER2/HER3 KD heterodimerization have not been reported. In the absence of such a molecule, we hypothesized that the combination of a HER2 TKI + PI3Ka inhibitor would block the increased oncogenicity caused by co-occurring HER2 and HER3 mutations. Indeed, the combination of neratinib and alpelisib strongly reduced growth and invasion of double-mutant cells. Similarly, the combination of HER2 and PI3Ka inhibitors has been suggested for HER2-amplified breast cancers harboring PIK3CA mutations (Hanker et al., 2013; Rexer et al., 2014). While initial clinical trials indicated that the combination of a pan-PI3K inhibitor with the HER2 TKI lapatinib resulted in significant toxicities (Guerin et al., 2017), a recent trial suggested that the combination of the HER2 antibody-drug conjugate T-DM1 and a more specific PI3Ka inhibitor is tolerable (Jain et al., 2018). Our results suggest that single-agent HER2 TKIs may not sufficiently block the growth of HER2-mutant tumors with co-occurring HER3 mutations. Therefore, clinical trials investigating the efficacy and safety of combining an HER2 TKI and PI3Ka inhibitor are warranted in cancers harboring cooccurring HER2/HER3 mutations.

5.4 Methods

5.4.1 Database searches

The Foundation Medicine database was queried for breast cancers harboring co-occurring mutations in ERBB2 and ERBB3 in January 2019. Breast cancers from METABRIC (n=2509), Broad (n=103), Sanger (n=100), TCGA (n=1108), INSERM Metastatic Breast Cancer (n=216), and the Metastatic Breast Cancer Project (n=237) were queried in April 2019 using www.cBioPortal.org (Cerami et al., 2012). Breast can-

cers from Project GENIE from Centers reporting alterations in ERBB2 and ERBB3 (n=8545; Centers = COLU, CRUK, DFCI, DUKE, MSK, PHS, UCSF, VHIO, VICC, and YALE) were queried in June 2019 using www.cBioPortal.org/GENIE(Consortium, 2017). All breast cancers with co-occurring ERBB2 and ERBB3 mutations were cross-referenced using at least two additional mutations in other genes to ensure that individual patients were not counted more than once.

5.4.2 Computational modeling

Structural modeling of proteins was carried out using the Rosetta 3.12 macromolecular modeling software package (Bender et al., 2016; Leman et al., 2020). The RosettaLigand application was used for molecular docking (Combs et al., 2013; Meiler and Baker, 2006). Molecular dynamics simulations were carried out using AMBER 18 (Case et al., 2018). Protein-protein interaction energy was obtained using the InterfaceAnalyzer mover in Rosetta. Protein-ligand interaction energy was estimated using MMPBSA.py (Miller et al., 2012). RMSD, atom-atom distances, and dihedrals angles were obtained using various applications: AmberTools (Case et al., 2018), CPPTRAJ (Roe and Cheatham, 2013), and Rosetta. We used the following forcefields / score functions for molecular modeling and simulation: AMBER ff14SB for proteins (Maier et al., 2015), generalized AMBER force field 2 (GAFF2) for ligands (neratinib), REF2015 for Rosetta kinase domain modeling, and Franklin 2019 for Rosetta HER2/HER3 near-full-length heterodimer modeling. Neratinib geometry optimization was performed with Gaussian 09 at the B3LYP/6-31G* level of theory. The electrostatic surface potential (ESP) was estimated with HF/6-31G* calculation. Partial charges generated with Gaussian 09 were fit to neratinib for MD simulations with the RESP procedure in AmberTools18 (Cornell et al., 1993). All structures were rendered with PyMOL 2.2. Graphs were generated with Matplotlib.

5.4.3 Structural modeling of the HER2-HER3 heterodimer

Modeling of the HER2/HER3 heterodimer was carried out in the Rosetta package (Song et al., 2013) utilizing multi-template comparative modeling (RosettaCM) with PDB structures 4RIW and 3PP0 as templates (Aertgeerts et al., 2011; Littlefield et al., 2014). HER3 was retained from 4RIW. The HER2 sequence was threaded on the receiver kinase EGFR structure from 4RIW during templated modeling, or was templated on the HER2 structure from 3PP0 superimposed on EGFR from 4RIW. In both instances, fragments from either structure were incorporated during RosettaCM refinement. Following the comparative modeling step, each structure underwent a single repeat of constrained FastRelax in the REF2015 score function. A total of 5000 structures were generated, and the top 20 best scoring structures were subjected to FastRelax with five repeats and constraints on starting coordinates. Constraints were ramped down during FastRelax. The best scoring complex was taken for subsequent analysis.

The near-full-length HER2/HER3 heterodimer was constructed with RosettaCM multi-template modeling. The HER2/HER3 KD heterodimer generated in the previous step, which included the juxtamembrane B (JMB) region, was used for the most of the intracellular component. C-terminal tails were excluded from modeling because they are primarily disordered. The transmembrane domain (TMD) and juxtamembrane A (JMA) regions were modeled based on the EGFR homodimer NMR structural ensemble in PDB ID 2M20. The HER2 extracellular domain (ECD) domains I – III were modeled from the HER2 crystallographic structure PDB ID 1N8Z. The HER3 ECD domains I – II were modeled from the EGFR crystallographic structure PDB ID 3NJP with fragments from the HER3 tethered structure PDB ID 1M6B. The PDB ID 1HAE NMR ensemble of Neuregulin 1 (NRG1) was superimposed with EGF from 3NJP prior to incorporation into the model of HER3 ECD. The ECD domain IV was modeled from 3NJP for both HER2 and HER3. Initial threaded models of each of these structures were combined with the Rosetta Domain Assembly application (Koehler Leman and Bonneau, 2018). Subsequently, the assembled structure underwent iterative rounds of all-atom minimization in the Franklin2019 score function with POPC implicit membrane and ramped constraints to start coordinates (weights successively lowered: 1.0, 0.5, 0.1, 0.0). The minimized structure was relaxed with constraints to start coordinates. Each domain (KD, JM, TM, and ECD) were separately and successively relaxed to produce 100 structures in each round, after which the best scoring structure was moved to the next round. The final structure was relaxed with constraints ramped down before being used in subsequent Rosetta mutational studies.

The fully inactivated $HER2^{WT}$ monomeric KD were generated with RosettaCM utilizing a structure of EGFR in the inactive state (PDB ID 3GT8) and refined with three independent 2.0 ms MD simulations. Structure snapshots were nominally collected every 20 ns from each trajectory and relaxed without constraints. The best scoring relaxed structure was taken to be the inactive HER2 conformation for steered MD and umbrella sampling simulations.

5.4.4 Molecular docking of HER2 protein and ligand (neratinib)

The initial structure of the inhibitor neratinib was downloaded from the PubChem database. The structures were then optimized using Gaussian 09 D.01 version at b3lyp/6-31G(d)* level. Electrostatic potential charges were calculated using Gaussian 09 and assigned using AmberTools. Small molecule conformers were generated with the BioChemical Library (BCL) conformer generator using default settings to create a maximum of 100 conformers (Mendenhall et al., 2020). Ligand (neratinib) docking was carried out using the RosettaLigand application in Rosetta 3.12 (Combs et al., 2013; DeLuca et al., 2015). The docking of ligands into proteins is divided into two phases: low resolution docking and high resolution docking. During the low-resolution docking phase, each ligand is allowed to explore the binding site in a 6.0 A radius. Rigid body

transformation is combined with ligand conformation swaps for 500 cycles of Monte Carlo Metropolis optimization. During the high-resolution docking phase, 6 cycles of side-chain rotamer and ligand conformer sampling were coupled with 0.2 $\AA$ in a Monte Carlo simulated annealing algorithm. 5000 docked protein-ligand complexes were generated. The interface score of the protein-ligand complex was calculated using the InterfaceAnalyzer mover in Rosetta 3.12 and the "ligand.wts" score weights. The root-mean-square deviation was computed using the lowest interface scored structure as the reference pose.

5.4.5 Classical MD simulations

Structures from the above modeling methods were used as an initial structure for further studies. The active and inactive reference frames of HER2 were set using previous studies and allowed to equilibrate based in our classical MD simulations. Each structure was solvated in a rectangular TIP3P box (12 $\AA$ buffer) neutralized with monovalent ions Cl and Na+ ions (Vega and Abascal, 2011). Solvent molecules were minimized with 2,000 steps of steepest gradient descent followed by 5,000 steps of conjugate gradient descent, while the protein/protein-ligand complex was restrained. The protein/protein-ligand complex was minimized in 2,000 steps of steepest gradient descent followed by 5,000 steps of conjugate gradient descent. Restraints were subsequently removed and the whole system underwent 2,000 steps of steepest gradient descent followed by 5,000 steps of conjugate gradient descent minimization. The system was slowly heated in NVT ensemble to 100K over 100 ps. The system was then heated in NPT ensemble at 1 bar from 100K to physiologic temperature (310K) over 500 ps. Equilibration was performed in NPT ensemble at 310K for 100 ns with a Monte Carlo barostat. The temperature was controlled using Langevin dynamics and a unique random seed was used for each simulation. SHAKE was implemented to constrain bonds involving hydrogen atoms. Periodic boundary conditions were applied and the particle mesh Ewald (PME) algorithm was adopted for the calculation of long-range electrostatic interactions with a cutoff distance of 10 A. Hydrogen mass repartitioning was employed to allow an integration time step of 4 fs.

5.4.6 Conformational free energy calculations

Potential of mean force (PMF) profiles for the active – inactive conformational transition in HER2 monomeric KD were obtained by performing constant velocity steered MD (SMD) and Umbrella sampling (US) simulations prior to free energy determination with the weighted histogram analysis method (WHAM) as implemented by Alan Grossfield (Grossfield,). SMD simulations were performed over 100 ns with a harmonic bias potential and spring constant of 500 kcal/mol/A 2. SMD simulations were performed in both directions (from the active to the inactive state and vice versa) using the Ca RMSD to the reference coordinates as the collective variable. A minimum of 250 windows were selected from each forward and backward simulation

with which to seed US simulations, such that each US simulation contained at least 500 windows to ensure overlap. A 2D harmonic restraining potential was applied to two CVs for the US simulations. CV1 (y-axis) was defined as the difference in the distance between R868(NE, CZ, NH1, NH2) – E770(OE1, OE2) and K753(NZ) – E770(OE1, OE2). CV2 (x-axis) was defined as the dihedral angle formed by the Ca atoms of the following residues: D863, F864, G865, and L866. A 2.0 kcal/mol/$\AA$ 2 spring constant was used for CV1, and a 10.0 kcal/mol/rad2 spring constant was used for CV2. At each umbrella center a 5 ns simulation was performed. The first 1 ns was used for equilibration, and the following 4 ns were used for analysis in WHAM. Lowest free energy pathway (LFEP) analysis completed with the LFEP package freely available from the Moradi Laboratory at the University of Arkansas.

5.4.7 Protein-ligand free energy calculations

Protein-ligand binding free energy calculations were performed with MM/GBSA implemented in the AmberTools18 MMPBSA.py (Miller et al., 2012). Trajectories were stripped of water and ions. Energies were computed with a surface tension of 0.0072 kcal/ mol/A 2 and salt concentration of 0.15 M. The non-polar contribution to the solvation free energy was approximated using the LCPO method (Weiser et al., 1999). Default radii assigned with Leap were kept for GBSA calculations. The enthalpic and solvation free energy contributions were computed every 100 ps. All calculations were completed from three independent trajectories and averaged.

5.4.8 Protein-protein interface energy

The protein-protein interface energy, or DG_dimerization, was determined using a modified version of the CartesianDDG protocol from Frenz et al. (Frenz et al., 2020). The best scoring $HER2^{WT}$/$HER3^{WT}$ KD heterodimer comparative model was transferred to the REF2015_Cartesian score function to an additional 20 rounds of FastRelax. The best scoring model from this subset was passed to the CartesianDDG application in Rosetta with interface mode enabled in order to generate optimized models for $HER2^{WT}$/$HER3^{WT}$, $HER2^{L755S}$/$HER3^{WT}$, $HER2^{V777L}$/$HER3^{WT}$, $HER2^{L869R}$/$HER3^{WT}$, $HER2^{WT}$/$HER3^{E928G}$, $HER2^{L755S}$/$HER3^{E928G}$, $HER2^{V777L}$/$HER3^{E928G}$, and $HER2^{L869R}$/$HER3^{E928G}$. The backbone degrees of freedom were set to i +/- 1 from the mutation site and 5 iterations were performed for each mutation. The all-atom attractive energy and solvation implicit energy score terms were given cutoffs of 9.0 $\AA$. Finally, an additional 100 structures were generated for each heterodimer KD complex by performing unrestrained Cartesian FastRelax beginning with the best scoring model by the dG_separated score term from the InterfaceAnalyzer mover (repacking both monomers after separation). Final binding affinity estimates for each complex are obtained by averaging the top 20 best structures by dG_separated from the final round of relax. Results are reported as mean +/- standard

error over those 20 models.

5.4.9 Plasmids

The Gateway Cloning system (Thermo Fisher Scientific) was used to generate pLX302-HER2 and pLX304-HER3 plasmids. The pDONR-223 vector encoding either $HER2^{WT}$ or $HER3^{WT}$ was subjected to site-directed mutagenesis (Genewiz) to generate HER2 or HER3 mutants. $HER2^{WT}$ and mutant plasmids were recombined into the lentiviral expression vector pLX-302 containing a C-terminal V5 epitope tag and puromycin resistance marker. $HER3^{WT}$ and mutant plasmids were recombined into pLX-304, also containing a C-terminal V5 tag, and blasticidin resistance marker. pFlag-CMV5.1 HER2 WT and HER3 WT ICDs were described previously (Hanker et al., 2017) and were subjected to site-directed mutagenesis (Genewiz) to generate mutants.

5.4.10 Transient transfections

Transient transfections were performed using Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's instructions. Co-transfection of pFlag-CMV5.1 HER2 and HER3 WT and mutant ICDs was performed as described (Red Brewer et al., 2013). siRNA transfections were performed using Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific) according to the manufacturer's instructions.

5.4.11 Lentiviral infections

Lentiviral supernatant was produced in early-passage 293FT cells by transfection with psPAX2 and pMD2.G packaging plasmids along with the appropriate pLX302-HER2 or pLX304-HER3 plasmid. Target cells or organoids were spin-infected the next day with viral supernatant in the presence of 8 mg/ml polybrene. Two d later, target cells/organoids were selected with puromycin (MCF10A: 2 mg/ml; OVCAR8: 0.7 mg/ml; MCF7: 0.5 mg/ml; SA493 organoids: 1 mg/ml) and/or 10 mg/ml blasticidin for at least 4 d. Stable cell lines were maintained in media containing puromycin and/or blasticidin.

5.4.12 Immunoprecipitation

If cells were pre-treated with antibodies (trastuzumab/pertuzumab), prior to lysis, cells were incubated with cold acid wash buffer (0.5 mol/L NaCl, 0.2 mol/L Na acetate, pH 3.0) for 6 min to remove bound antibodies. Monolayers were then washed 3 times with ice-cold PBS. Cell lysates were harvested using ice ND lysis buffer [1% Triton X100, 20 mM Tris HCl, 150 mM NaCl, supplemented with 1X protease inhibitor (Roche) and phosphatase inhibitor (Roche) cocktails] and rotated at 4°C for 1 h. Lysates were then clarified by spinning at 10,000 3 g at 4°C for 15 min. Protein concentrations were measured using BCA standard curves (Pierce). Four-eight mL of HER2 Ab-17 antibody (Thermo Fisher Scientific) was added to 500-1000 mg

protein lysate and rotated at 4°C overnight. IP was carried out using the Invitrogen Dynabeads Protein G Immunoprecipitation Kit (10007D) as directed. Lysates were next subjected to SDS-PAGE and immunoblot analysis. Each immunoprecipitation experiment was performed a minimum of two times.

5.4.13 Proximity ligation assay

MCF10A cells (5 x 104 cells/well) were seeded in 8-well chamber slides (Lab-Tek, 177445) in triplicate and incubated in EGF/insulinfree media + 1% CSS overnight. PLA was performed with Duolink In Situ Red Starter Kit Mouse/Rabbit (Sigma) using mouse antiHER2 (Thermo Fisher Scientific; Cat MS-730-P1-A) and rabbit anti-HER3 (Cell Signaling Technologies; Cat 12708) antibodies according to the manufacturer's protocol and then imaged with a DMi8 inverted microscope (Leica). The number of PLA foci per cell was quantified using ImageJ as described (Prado Martins et al., 2018). A minimum of 7 images per sample were analyzed.

5.4.14 Western blot analysis

Prior to lysing, organoids were dissociated into single cell suspension by mechanical shearing and enzymatic digestion using TrypLE express (Gibco, 12604021). Adherent cells or organoid cell pellets were washed with ice-cold PBS and lysed with RIPA buffer (Sigma) supplemented with 1X protease inhibitor (Roche) and phosphatase inhibitor (Roche) cocktails. Lysates were centrifuged at 13,500 rpm for 15 min. Protein concentrations in supernatants were quantified using BCA protein assay kit (Pierce). 20-40 mg of total protein was fractionated on bis-tris 4-12% gradient gels (NuPAGE) and transferred to nitrocellulose membranes (BioRad). Membranes were blocked with 5% non-fat dry milk/TBST at room-temperature for 1 h, followed by overnight incubation with primary antibodies of interest at 4C in 5% BSA/TBST. All antibodies were purchased from Cell Signaling – P-HER2 Y1221/2 (2243; 1:500), HER2 (2242; 1:1000), P-HER3 Y1197 (4561; 1:500), P-HER3 Y1289 (4791; 1:500), P-HER3 Y1197, HER3 (12708; 1:1000), P-AKT S473 (9271; 1:500), P-AKT T308 (13038; 1:500), P-S6 S235/6 (2211; 1:1000), PS6 S240/4 (2215; 1:1000), P-ERK T202/Y204 (9101; 1:1000), and b-actin (4970; 1:1000). Membranes were cut horizontally to probe with multiple antibodies. In some cases, P-Akt S473, P-Erk, and P-S6 S240/244 antibodies were combined during primary incubation. Nitrocellulose membranes were washed and incubated with HRP-conjugated a-rabbit or a-mouse secondary antibodies for 1 h at room temperature. Protein bands were detected with an enhanced chemiluminescence substrate (Perkin Elmer) using the ChemiDoc Imaging System (Bio-Rad). Immunoblots were quantified using ImageJ.

5.4.15 Flow cytometry

HER-2 cell surface staining was performed with the trastuzumab antibody. MCF10A stable cells (8 x 105) were incubated with 0.2 mg/ml trastuzumab for 30 min at 4°C. Cells were washed in FACS buffer (Thermo Scientific) then incubated with an Alexa Fluor 647-conjugated goat anti-human IgG secondary antibody (Thermo Scientific; 1 mg/ml) for 30 min at 4°C. After 2 additional washes, the cells were analyzed on an LSR Fortessa flow cytometer (BD Biosciences). Ten thousand cellular events were analyzed per sample. Data were analyzed using FlowJo software (BD Biosciences).

5.4.16 Organoid establishment and culture

Fresh/frozen tumor chunks from SA493 ($HER2^{S310F}$) PDXs were rinsed twice with 10 ml AdDF+++ media (advanced DMEM/F12 containing 1X Glutamax, 10 mM HEPES and antibiotics) and minced into 1-2 mm pieces. 10 ml dissociation media (1:1 vol/vol F12, DMEM supplemented with 2% w/v bovine serum albumin, 300 U/ml collagenase, 100 U/ml hyaluronidase, 10 ng/ml epidermal growth factor (EGF), 1 mg/ml insulin, and 0.5 mg/ml hydrocortisone) was added to tumor fragments and incubated for 2 hr at 37°C with constant shaking at 275 rpm. Dissociated tumor fragments were centrifuged at 1200 rpm for 5 min and subjected to RBC lysis as per manufacturer's protocol (BD Biosciences), if the cell pellet was visibly red. Tumor fragments were further dissociated by adding 3 ml pre-warmed trypsin and incubating in a 37C bead bath for 5-7 min. 6 ml neutralization solution (2% FBS in PBS) was added and centrifuged at 1200 rpm for 5 min. Tumor pellets were then treated with the Dispase/DNAse cocktail for 5-7 min at 37°C, and neutralized and centrifuged as above. Tumor cell suspension was subjected to magnetic separation of CD298+ human cells (biotin-conjugated a-CD298 antibody, Miltenyi Biotec, 130-101-292) to eliminate potential mouse cell contamination, using EasySep human biotin positive selection kit II (STEMCELL technologies 17663). The cell pellet was resuspended in appropriate volume of cold BME and 40 ml of cell suspension was added to the center of each well of a 24-well plate and allowed to solidify by placing in a 37°C incubator for 20 min. 500 ml organoid medium (DMEM/F12 containing 250 ng/ml R-Spondin 3, 5 nM Neuregulin 1, 5 ng/ml FGF7, 20 ng/ml FGF10, 5ng/ml EGF, 100 ng/ml Noggin, 500 nM A83-01, 5 mM Y-27632, 500 nM SB202190, 1X B27 supplement, 1.25 mM N-Acetylcysteine, 5 mM Nicotinamide, 1X GlutaMax, 10 mM Hepes, 50 mg/ml primocin, and 100 U/ml penicillin/100 mg/ml streptomycin) was added to each well and the plate was returned to a 37°C incubator maintained at 2% O2 level. For viability assays, established organoids were dissociated into single cell suspension by mechanical shearing and enzymatic digestion using TrypLE express (Gibco, 12604021). Dissociated cells were resuspended in 100 ml of cold organoid media containing 5% BME and 1000 cells/well were seeded into BME-coated 96-well plate in organoid media lacking EGF and NRG1. The next day, organoid cultures were treated with drugs and the effect on viability was assessed

6 d later using CellTiter-Glo 3D viability assay kit (Promega G9681). Organoids were photographed using a Leica DMi1 inverted microscope.

5.4.17 Sanger sequencing of ERBB2 and ERBB3

RNA was isolated from CW2 cells using the Maxwell RSC simplyRNA Cells Kit (Promega) on the Maxwell RSC Instrument (Promega). RNA was isolated from SA493 organoids using the Qiagen RNeasy Micro Kit. Reverse transcription was performed using the iScript cDNA Synthesis Kit (Bio-Rad). The appropriate regions of ERBB2 and ERBB3 were PCR-amplified using the following primers: 5'GCCTGCCTCCACTTCAACCA (ERBB2_foward; S310F), 5' GTAACTGCCCTCACCTCTCG (ERBB2_reverse; S310F), 5' GTGAAGGTGCTTGGATCTGG (ERBB2_foward; L755S), 5' ATCTGCATGGTACTCTGTCT (ERBB2_reverse; L755S), 5' TGAGGCGATACTTGGAACGG (ERBB3_forward), and 5'AGGTTGGGCGAATGTTCTCA (ERBB3 reverse). Sanger sequencing for ERBB2S310F, ERBB2L755S, and ERBB3 was performed using the 5' CATCTGTGAGCTGCACTGCC, 5'GTTGGGACTCTTGACCAGCA, and 5'GTGCATAGAAACCTGGCTGC sequencing primers, respectively.

5.4.18 Quantitative RT-PCR

Total RNA was isolated using the Maxwell RSC simplyRNA Cells Kit (Promega) on the Maxwell RSC Instrument (Promega). cDNA was synthesized using the iScript cDNA synthesis Kit (Bio-Rad) and then subjected to qPCR using PowerUp SYBR Green Master Mix (Thermo Fisher Scientific) and Qiagen RT2 qPCR primer assays for human ERBB2, ERBB3, and YWHAZ (housekeeping control). To specifically detect ERBB22264T>C (L755S), the following qPCR primers were used: 5'CAGTGGCCATCAACGTGTC (forward) and 5'TACACCAGTTCAGCAGGTCCT (reverse). qPCR was performed using the QuantStudio3 Real-Time PCR System (Thermo Fisher Scientific).

5.4.19 Cell viability assay and IC50 estimation

Cell viability was determined using the Cell Titer Glo assay (Promega) according to the manufacturer's instructions. Briefly, singe-cell suspensions were generated by straining trypsinized cells through a 40mm cell strainer (Fisher Scientific). 500-1000 cells per well were plated in 96-well white clear-bottom plates in quadruplicate. Cells were treated with 10 concentrations of inhibitor or vehicle alone at a final volume of 150 mL per well. After 6 d of treatment, 25 mL of Cell Titer Glo was added to each well. Plates were shaken for 15 min, and bioluminescence was determined using the GloMax Discover Microplate Reader (Promega). Blank-corrected bioluminescence values were normalized to DMSO-treated wells and normalized values were plotted in GraphPad Prism using non-linear regression fit to normalized data with a variable slope (four

parameters). IC50 values were calculated by GraphPad Prism at 50% inhibition.

5.4.20 Cell proliferation assay

CW2 cells were transfected with Control or HER3 siRNA in triplicate. Four d after transfection, cells were trypsinized and counted with a Z2 Coulter Counter Analyzer (Beckman coulter).

5.4.21 Three-dimensional morphogenesis assay

Cells were seeded on growth factor–reduced Matrigel (BD Biosciences) in 48-well plates following published protocols (Debnath et al., 2003). Inhibitors were added to the medium at the time of cell seeding. Fresh media and inhibitors were replenished every 3d. Following 7-10 d, colonies were stained with 5 mg/ml MTT for 20 min. Plates were scanned and colonies measuring R100 mm were counted using GelCount software (Oxford Optronix). Colonies were photographed using a Leica DMi1 inverted microscope.

5.4.22 Cell invasion assay

Transwell invasion assays were performed using BioCoat Growth Factor Reduced Matrigel Invasion Chambers (Corning) according to the manufacturer's instructions. Briefly, MCF10A cells were seeded at 100,000 cells/well in serum-free DMEM/F12 media. DMEM/F12 media containing 5% FBS was added to the bottom chamber as a chemoattractant. The cells were incubated under the desired conditions and 22 h later, cells that invaded to the underside of the membrane were stained with 0.5% crystal violet. Transwells were photographed using a Leica DMi1 inverted microscope. Brightfield images were quantified using ImageJ software. Images were converted to RGB stack. The green channel was thresholded and filtered (3 pixels) to remove the pores. The total thresholded area was measured.

5.4.23 Xenograft Studies

CW2 cells were re-suspended in serum-free RPMI and Growth Factor-Reduced Matrigel (1:1 ratio) and injected subcutaneously into the right flank of 4-6 week old female athymic nu/nu mice (Envigo). When the average tumor volume reached $\tilde{2}$00 mm3, mice received daily doses of vehicle (0.5% Methylcellulose + 0.4% Tween 80, orogastric gavage), neratinib (40 mg/kg; orogastric gavage), alpelisib (30 mg/kg; orogastric gavage), or neratinib + alpelisib. In our previous studies, we have found neratinib to cause anorexia and moderate body weight loss. To avoid these toxicities, all mice were prophylactically supplemented with DietGel 76A (Clear H2O) in addition to regular chow. Tumor diameters were measured twice weekly using calipers and tumor volumes were calculated using the formula: volume = width^2 x length/2.

5.4.24 Quantification and statistical analysis

Statistical analysis was performed using GraphPad Prism 8.1.2. For analyses involving multiple comparisons, one-way or two-way (for grouped bar graphs) ANOVA with Bonferroni posthoc test was used. Otherwise student's t-test was used. Bar graphs show mean +/- S.E.M. The neratinib/alpelisib combination index was calculated using the Chou-Talalay test (Chou, 2010).

CHAPTER 6

BCL::MolAlign: Three-Dimensional Small Molecule Alignment for Pharmacophore Mapping

This chapter is taken from Brown, B. P.; Mendenhall, J.; Meiler, J. J. Chem. Inf. Model. 2019, 59 (2), 689–70138.

6.1 Introduction

Small molecule flexible alignment is the process of organizing 3D molecular structures in space according to their similarities. It is a necessary step in a number of computer-aided drug discovery (CADD) strategies that utilize 3D structural information to evaluate putative ligands (Wolber et al., 2008). Ligand alignment is necessary because the protein-bound ligand pose is distinct from the pose adopted by the ligand in free solution (Vieth et al., 1998; Perola and Charifson, 2004; Hao et al., 2007). Bioactive ligand conformations result not just from low-energy ligand conformational selection, but also from protein conformational accessibility (Seo et al., 2014; Greives and Zhou, 2014). Consequently, the binding conformation of the ligand cannot be reliably determined by minimizing ligand internal strain alone if the ligand can adopt multiple conformations that have comparable energy.

Determination of the most likely ligand binding-pose is a critical component of ligand- and structure-based drug discovery. One of the most extensively utilized and actively developed methods in CADD is pharmacophore modeling (Yang, 2010). Recent innovation has led to the development of interactive software for building pharmacophores and designing lead compounds from them (Vlachakis et al., 2015; Beccari et al., 2013). Among the most significant challenges in pharmacophore modeling is obtaining an accurate and informative molecular alignment (Yang, 2010). Structure-based methods are also enhanced by effective molecular alignment. Effective protein-ligand docking usually requires a priori knowledge of an approximate binding mode (Leelananda and Lindert, 2016; Sliwoski et al., 2014; Hecker et al., 2002; Kubinyi, 2003). Despite significant advances in the field (Cleves and Jain, 2018; Chan, 2017; Roy and Skolnick, 2015; Urniaz and Jozwiak, 2013; Thormann et al., 2012; Sastry et al., 2011; Tosco et al., 2011; Korb et al., 2010; Heifets and Lilien, 2010; Jain, 2007; Richmond et al., 2006; Wildman and Crippen, 2001), small molecule flexible alignment remains a challenging problem. Here, we present a novel small-molecule flexible alignment algorithm in the BioChemical Library (BCL) molecular modeling suite called BCL::MolAlign.

A successful alignment algorithm must provide: (1) Efficient sampling of each molecule's conformational space, (2) efficient sampling of possible alignments, and (3) scoring aligned poses according to their fit. There are generally two strategies employed to account for ligand flexibility during the search procedure: (1) Rigid-

body alignment with an ensemble of molecule conformers, or (2) bond angle sampling as a discrete step during alignment7. Methods which rely exclusively on pre-generated conformers (e.g. LIGSIFT, ROCS, Shapelets, PL-PatchSurfer) are limited in their predictive potential by the initial conformers produced (Roy and Skolnick, 2015; McGaughey et al., 2007; Tawa et al., 2009; Proschak et al., 2007; Tervo et al., 2005; Cheeseright et al., 2006; Shin et al., 2015). Several other approaches, such as FlexS (Andrews and Cramer, 2000) or the flexible alignment software available through Chemical Computing Group's MOE (Labute et al., 2001; Chan and Labute, 2010), account for ligand flexibility by including torsional sampling during the alignment procedure. These algorithms must simultaneously enforce rules minimizing ligand internal strain against rules maximizing alignment score. This can result in unrealistic ligand poses in cases where the molecules being compared are of substantially different size or shape (Labute et al., 2001).

To address deficiencies in conformational sampling, we have implemented a unique combination of both of the above approaches. We first utilize BCL::Conf to generate an ensemble of conformers for one or both molecules. The difficulty in applying pre-generated conformations for molecular alignment is generating native-like, physically realistic conformations. BCL::Conf combines a CSD-derived rotamer library with a conformer scoring function based on dihedral rotamer propensity and atomic clashes to rate the likelihood of a given conformer. With this scoring scheme, BCL::Conf is able to recover more native-like conformers than other widely used conformer generation protocols (Kothiwale et al., 2015). We subsequently apply limited on-the-fly flexible refinement of the target conformer during pose sampling. On-the-fly conformational changes that do not pass the BCL::Conf clash score are rejected.

An additional challenge is in developing a robust search algorithm to navigate the shared conformational space (co-space) of the molecules being aligned. The majority of programs employ a deterministic algorithm based on maximum overlap of molecular volume (Roy and Skolnick, 2015; McGaughey et al., 2007; Tawa et al., 2009). While rapid, such an approach necessarily becomes less effective as the number of rotatable bonds (and correspondingly, the non-degenerate conformations) of the target molecule increases. We address this deficiency by utilizing multi-trajectory Monte Carlo Metropolis (MCM) sampling to overlay nearby substructures of the molecules. Our method allows rapid convergence on the co-space of the molecules while maintaining dynamic conformational sampling. Moreover, BCL::MolAlign may optionally superimpose molecules based on maximal common substructures defined by specific atom and bond type features.

Finally, a scoring metric is needed that is capable of ranking molecule superimpositions based on the degree to which chemically similar functional groups are best superimposed. Many algorithms implement a Tanimoto coefficient to grade chemical and/or shape similarity (Roy and Skolnick, 2015; McGaughey et al., 2007; Tawa et al., 2009). Strict Tanimoto comparisons are incapable of grading alignments when the

molecules being compared are of sufficiently different sizes. This prohibits accurate alignment and ranking of derivatives to substructure scaffolds. Many methods are based on Gaussian overlap, where a Gaussian decay is applied to each property and the score is simply the 3D-spatial integral of the overlap, often computed solely at the centers of each atom (Roy and Skolnick, 2015; Vainio et al., 2009). This approach suffers from the offset problem – if the properties are continuous, such as van-der Waals volume, then the optimal alignment of two atoms very different in size will be offset. Additionally, Gaussian based methods typically define a single length scale for each property, which is arbitrary and inappropriate for binding pockets of different levels of flexibility (Vainio et al., 2009). An alternative approach is to generate a comparison function from weighted linear combinations of chemical properties (Chan and Labute, 2010). We took the latter approach; our scoring function is computed by summing the weighted property-distance between nearest-neighbor atoms of the molecules being aligned. Our method has the added advantage that atoms in one molecule that have no corresponding partner in the other molecule do not influence the search procedure.

6.2 Results

6.2.1 BCL::MolAlign uses a three-tiered Monte Carlo Metropolis protocol to identify optimal superimpositions for two molecules

BCL::MolAlign perturbations are implemented primarily through a Monte Carlo Metropolis (MCM) search procedure (Table 6.1).

An overview of the algorithm is presented in Figure 6.1. Briefly, at least one MC trajectory is performed for each alignment with the option to specify additional independent trajectories. Each trajectory will perform three tiers of optimization (Figure 6.1). In the first tier, pre-generated conformer pairs (one from each molecule) undergo limited optimization to remove the lowest scoring 25% of conformer pairs. The total number of conformer pairs tested is a user-specified quantity. Tier two iteratively refines the best alignments and removes the lowest scoring user-specified fraction after each iteration. Tier three performs a final optimization of the top N user-specified pairs from round two. BCL::MolAlign can align a single target molecule against another ligand in a known binding pose (herein referred to as the "scaffold" ligand), or it can independently move both molecules in a pair to optimize their alignment.

Each step of the MCM is scored. If the score is the best that has been sampled so far, or if it is improved over the previously accepted step, then that step is automatically accepted. If the score is not improved then there is a probability that it will be accepted dependent on the magnitude of the score difference and the temperature (Figure 6.1). The temperature automatically adjusts to satisfy user-specified acceptance ratios over the course of the simulation (Karakas et al., 2012).

At the beginning of each alignment, BCL::Conf will attempt to generate a user-specified number of con-

Move	Use	Description
BondAlign	Rigid or Flexible	Superimpose an individual bond from two nearest-neighbor atoms in each molecule
BondAlign2	Rigid or Flexible	Superimpose two bonds from two nearest-neighbor atoms in each molecule
MatchAtomNeighbors	Rigid or Flexible	Superimpose all matched atom pairs within the maximum distance threshold.
BondSwap	Rigid or Flexible	Transform the molecule such that the position coordinates of a random bonded atom pair are swapped with the position coordinates of a second random bonded atom pair
RotateSmall	Rigid or Flexible	Randomly rotate the molecule 0 – 5 degrees about a randomly-selected axis
RotateLarge	Rigid or Flexible	Randomly rotate the molecule 0 – 180 degrees about a randomly-selected axis
BondRotate	Flexible only	Randomly rotate non-amide, non-ring, outermost single bond between heavy atoms that form dihedral angles with adjacent heavy atoms
ConformerSwap	Flexible only	Swap a current conformer for another in the library

Table 6.1: **Summary of sampling strategies employed in BCL::MolAlign.**

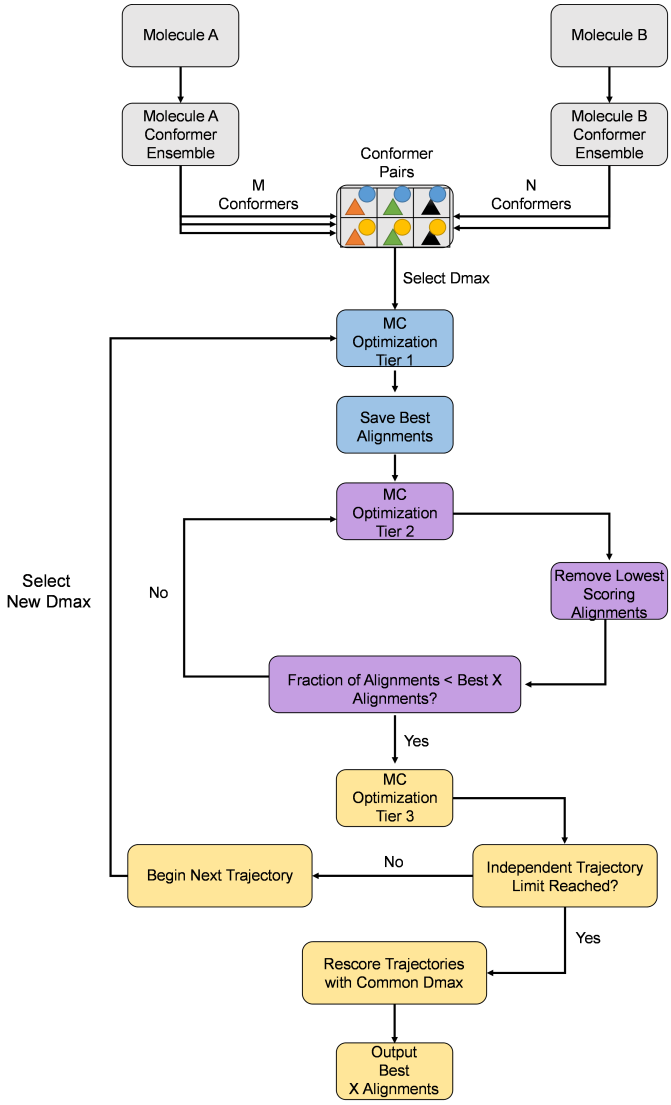


Figure 6.1: Outline of the BCL::MolAlign flexible alignment algorithm. Rigid alignment is equivalent to a single tier of MCM optimization with a single conformation each for Molecule A and Molecule B. MC moves alter the current Molecule A or B during each optimization tier. The same moves are used in each tier, but number of steps differ in each tier.

formations, or a default number of 100 unique conformations, for each molecule for which flexibility is allowed. Subsequently, conformers of the two molecules will be randomly paired until the number of conformer pairs is equal to the minimum of the total number of possible pairs and a user-specified conformer pair number (the default conformer pair number is equal to 100 pairs). For example, if BCL::Conf generates 50 conformers of each of the two molecules being aligned, then there are 2500 possible conformer pairs. With the default settings, 100 conformer pairs would be randomly selected as starting points for alignment. An MCM ConformerSwap mover is implemented to allow access to the other 2400 possible conformer pairs during the alignment. Alternatively, if each molecule has only one conformation, then only one conformation pair would be selected as a starting point because the total number of possible pairs is less than the default conformer pair number of 100.

Conformational sampling is incorporated into the search procedure through a combination of pre-generated conformer swapping and on-the-fly bond rotation. Specifically, we either swap one conformer for a separate conformer from those generated at the beginning of the alignment with BCL::Conf (ConformerSwap), or rotate particular bonds (BondRotate; Table 6.1). ConformerSwap randomly selects a conformer from the entire conformational ensemble of one of the molecules in the pair. The coordinates of that conformer in 3D real space are then transformed to minimize the RMSD to the original conformation of the same molecule.

BondRotate rotates non-conjugated, non-ring, single bonds between heavy atoms that form dihedral angles with adjacent heavy atoms. To ensure that the bond rotation yields an energetically favorable conformation, we first obtain a set of allowed rotations for each dihedral from BCL::Conf's rotamer library. Initially we observed that this move was very rarely accepted when it was applied to bonds near the core of the molecule, presumably because altering a dihedral near the core of the molecule often perturbs the entire conformation. Likewise, we restricted BondRotate to only work on the outermost heavy-atom dihedral angles in the molecule. The purpose of BondRotate is to allow refinement of otherwise well-aligned conformers when the probability of substituting the correct conformer is prohibitively low or null due to the necessarily incomplete coverage of conformational space. If BondRotate results in a molecule conformation which does not satisfy the BCL:Conf atom clash score (Kothiwale et al., 2015) then the move is rejected prior to scoring and an alternative MCM move is attempted.

BCL::Conf, and by extension BCL::MolAlign, does not perform explicit calculations of conformer internal energy, and instead relies on statistical potentials. While conformers with higher internal strain can potentially be sampled, it is also possible for protein-bound ligands to exhibit conformers of higher internal energy relative to the solution state (Perola and Charifson, 2004; Hao et al., 2007). If additional restrictions on acceptable conformers are desired, conformation sampling can easily be turned off, and externally generated conformers can be used as the input for separate rigid alignment runs.

6.2.2 BCL::MolAlign iteratively samples alignments through superimposition of bonded atoms

In addition to conformational changes, BCL::MolAlign samples possible alignments through multiple movers, or sampling functions, implemented in the MC protocol. The most intuitive perturbations for both rigid and flexible alignment implemented in BCL::MolAlign are rotation and translation of a whole molecule. Translate1 translates molecules between 0-1 $\AA$ (uniformly distributed) from their starting positions, in a randomly chosen direction. RotateSmall rotates molecules between 0-5°, uniformly distributed on the unit sphere within these bounds, from their starting conformations. RotateLarge rotates molecules randomly between 0-180° (Kuffner, 2004). BCL::MolAlign also utilizes a series of moves designed to superimpose the coordinates of nearest-neighbor atoms (BondAlign, BondAlign2, and MatchAtomNeighbors) without explicitly comparing common substructures. BondAlign, BondAlign2, and MatchAtomNeighbors provide progressively higher resolution sampling of the local alignment space.

Consider two molecules, designated A and B. The BondAlign mover identifies in A the heavy atom that is nearest in Cartesian space to a randomly-chosen heavy atom in B, irrespective of their atom types. BondAlign then superimposes a randomly-chosen bond from the selected atoms of A and B (Figure 6.2A).

Similarly, BondAlign2 superimposes two randomly-chosen bonds of a randomly-selected atom (S) in A with two randomly-selected bonds from the closest atom in B to S. Only atoms with two or more bonds are considered for this step. (Figure 6.2B).

The MatchAtomNeighbors mover computes all mutually nearest atom pairs between A and B within a maximum distance threshold (see subsection Variable distance cutoffs dictate which atom pairs are included in alignment scoring). Subsequently, A is transformed such that the total mean square distance between the mutually nearest atoms in A and B is minimized (Figure 6.2C).

BondSwap differs from the previous three movers in that it is not based on nearest-neighbor atoms between the two molecules being aligned. The BondSwap mover randomly selects two unique bonds between heavy atoms within A. The molecule is rotated and translated such that the position of the first bond becomes the position of the second bond, or vice versa (Figure 6.2D).

The probability that a particular mover is selected is proportional to the total amount that each mover improved the scores on the Astra-Zeneca overlay set when all movers were used with equal probability.

6.2.3 Variable distance cutoffs dictate which atom pairs are included in alignment scoring

The scoring system was inspired by previous work in our lab, which used Euclidean distance combined with a property value as an additional dimension to evaluate docked conformations of mGluR allosteric modulators (Gregory et al., 2014). In the present study, we expanded that score function to compute the weighted property distances between atom pairs.

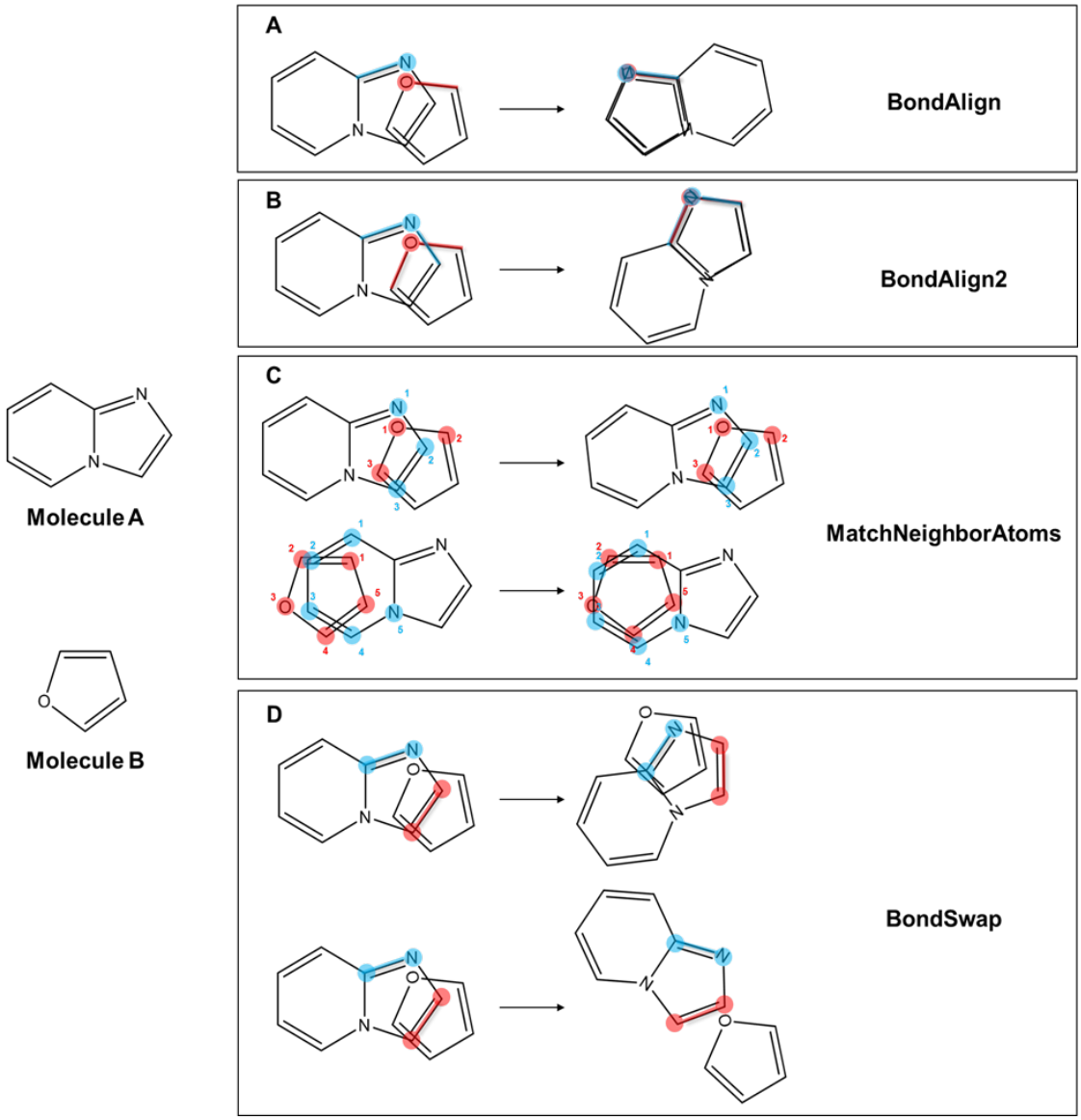


Figure 6.2: Schematic of sampling strategies implemented in BCL::MolAlign. From a given starting alignment on the left side of the arrow, the resulting alignment following each operation is depicted on the right side of the arrow. Once atoms and bonds have been chosen, BondAlign (A), BondAlign2 (B), and MatchAtomNeighbors (C) each have one possible outcome. BondSwap (D) has an equal probability of sampling two possible outcomes. Highlighted segments correspond to the chosen atoms and bonds for alignment. Atom numberings in MatchNeighborAtoms correspond to mutually matched pairs between molecules A and B.

For a given alignment of molecules M_a and M_b, $W(M_{b,i,j})$ is the matching weight of the j-th atom of molecule M_b on the i-th atom of molecule M_a, and is defined by:

$$W(M_b,i,j)=\begin{cases}\frac{1}{2}\left(cos\left(\pi\frac{D(M_{a,i},M_{b,j})}{D_{max}}\right)+1\right) & D(M_{a,i},M_{b,j})\le D_{max}\\ 0 & D(M_{a,i},M_{b,j})>D_{max}\end{cases} \tag{6.1}$$

where $D(M_{(a,i)},M_{(b,j)})$ is the distance of the i-th atom in molecule M_a from the j-th atom in molecule M_b. D_{max} is the maximum distance cutoff determining whether or not two atoms are paired (Figure 6.3). Similarly, we compute the matching weight of the j-th atom of M_a on the i-th atom of M_b as $W(M_{a,i,j})$.

For the vast majority of atoms, there is a simple one-to-one matching between these atom pairs based on distance in our alignments. This enables a simplistic comparison of the properties on the associated atoms without any need for weighting relative contributions from other nearby atoms. However, our scoring function maintains the capacity to handle the cases where an atom straddles a covalently bonded atom pair in the other molecule (Figures 2D, 3).

D_{max} is randomly selected in each independent MCM trajectory from a user-defined range. In this benchmark, each alignment was run with five independent trajectories each of which sampled a D_{max} between 0.7 and 1.2 $\AA$. 0.7 $\AA$, the covalent radius of quaternary carbon, was chosen as the lower bound to allow a single carbon atom to straddle anywhere along a C-C bond, while effectively only matching to the nearer of the two C-C atoms. The upper cutoff of 1.2 $\AA$ was nominally chosen as the smallest covalent diameter of any common heavy atom type, alkyl-carbon (0.6 $\AA$ radius), to prevent smearing caused by neighboring heavy atoms. To allow comparison between the independent trajectories, the overall best alignments from each trajectory are re-scored at a maximum atom distance 1.0 $\AA$ to determine which trajectory yielded the best alignment.

Next, we compute the weighted property average of property p in molecule M_b at the coordinates of the i-th atom in M_a, denoted by $\overline{PN(M_{b,p,i})}$:

$$\overline{PN(M_b,p,i)}=\begin{cases}\frac{\sum_{j=1}^{N_b}p_{b,j}W(M_b,i,j)}{\sum_{j=1}^{N_b}W(M_b,i,j)} & if\ \sum_{j=1}^{N_b}W(M_b,i,j)\neq 0\\ 0 & if\ \sum_{j=1}^{N_b}W(M_b,i,j)=0\end{cases} \tag{6.2}$$

The property square norm for property p, computed for molecules M_a and M_b is the squared L2 norm between

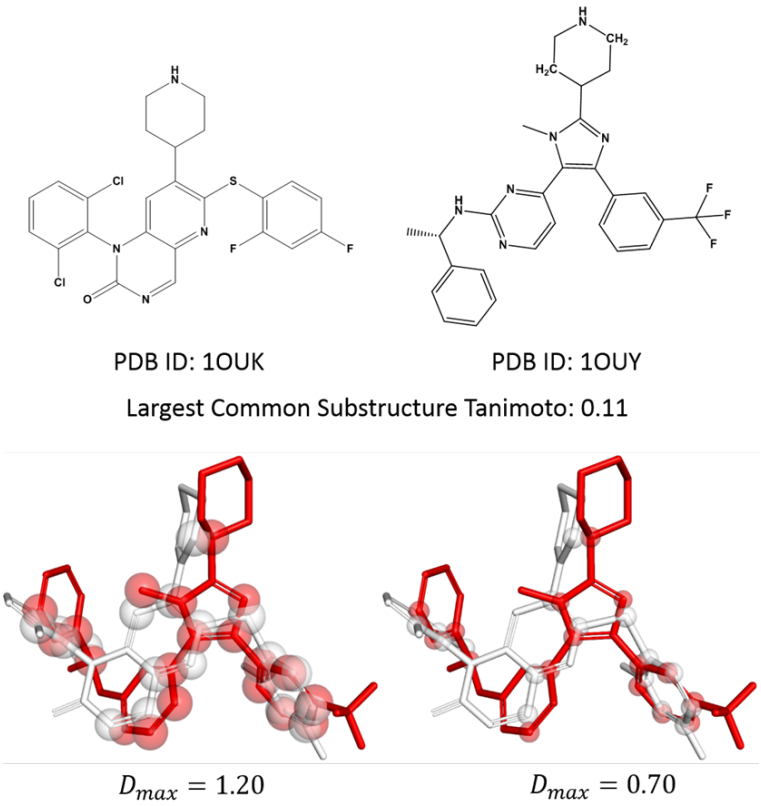


Figure 6.3: Rigid alignment of P38 inhibitors from PDB IDs 1OUK and 1OUY illustrate atom pairing at variable maximum atom distances. The 2D representations of the 1OUK and 1OUY ligands. The 3D representations depict the native pose of 1OUK rigidly aligned to the native pose of 1OUY. Spheres illustrate heavy atoms separated from a heavy atom in the opposite molecule by less than the specified maximum atom distance D_{max}. Sphere radii correspond to half of the indicated maximum atom distance. Red and white overlapping spheres are considered matched atoms.

a property of atoms of molecule A and the corresponding matched atoms in B:

$$PSN(M_a, M_b, p) = \sum_{i=1}^{N_{atoms,a}} \left(p_{M_{a,i}} - \overline{PN(M_b, p, i)}\right)^2 \tag{6.3}$$

where $p_{(M_{(a,i)})}$ represents the value of property p for the i-th atom of molecule M_a. This norm is asymmetric with respect to A and B, reflecting the notion that A molecule may very well cover the pharmacophore of B, while the converse is untrue. For some applications (i.e. clustering), a symmetric measure of dissimilarity is desired which is ideally normalized to 0-1. Likewise, we define the normalized property distance, $P_{Norm}(M_a, M_b, p)$, between molecules M_a and M_b:

$$P_{Norm}(M_a, M_b, p) = \sqrt{\frac{PSN(M_a, M_b, p) + PSN(M_b, M_a, p)}{P_{M_a}^2 + P_{M_b}^2}} \tag{6.4}$$

where $P_{(M_a)}^2$ is the sum of property value squares.

The total property distance between molecules M_a and M_b is determined by computing the weighted sum of the normalized property distances for all specified properties as (eq. 1):

$$D(M_a, M_b) = \frac{\sum_{p=1}^{Np} P_{Norm}(M_a, M_b, p) w_p}{\sum_{p=1}^{Np} w_p} \tag{6.5}$$

where w_p is the weight of property p. Property weights were obtained as previously described by computing the inverse standard deviation of each property's occurrence across a sample library of drug-like small molecules (Gregory et al., 2014; Butkiewicz et al., 2013), so as to nominally give each property equal weight or influence over the results.

We noted that the size of the core subset of atoms in a molecule responsible for conferring bioactivity may vary dramatically between targets. However, approximately 80% of the experimentally-determined pharmacophores available in the AstraZeneca Overlays Validation Test Set have at least 60% heavy atom overlap in their natively bound poses given a 1.0 $\AA$ max atom distance (Giangreco et al., 2013). Consequently, the final property distance score (PDS) is computed such that alignments with less than 60% of their total

heavy atoms matching are penalized:

$$PDS(M_a, M_b) = D(M_a, M_b) + PEN(m) \tag{6.6}$$

where penalty as a function of the total fraction of atoms matched, $PEN(m)$, is defined

$$PEN(m) = \begin{cases} C\left(\frac{0.6 - m}{0.6}\right)^2 & m < 0.6 \\ 0 & m \geq 0.6 \end{cases} \tag{6.7}$$

for a user-specified base mismatch penalty, C, and the ratio of paired-to-unpaired atoms, m. For the purposes of our benchmark, we nominally took C to be 2.0. The single alignment that minimized eq. 6 was taken to be the final alignment. The BCL allows customizable implementation of molecule and atom descriptors for a multitude of tasks (Butkiewicz et al., 2013).

6.2.4 BCL::MolAlign improves recovery of crystallographically-determined ligand binding poses

To evaluate the efficacy of our method in recovering native ligand binding-poses, we used a previously published benchmark set of small molecule inhibitors for six protein targets: CDK2, HIV, P38, ESR1, Trypsin, and Rhinovirus (Chan and Labute, 2010; Chen et al., 2006) (Table 6.2). For two of the datasets, P38 and ESR1, we also evaluated BCL::MolAlign on two previously distinguished pharmacophores (Chan and Labute, 2010; Chen et al., 2006), which yielded an additional four test cases. For each of the datasets, an NxN pairwise alignment of every molecule was performed. Rigid alignments were initiated by centering the native bound conformers of each ligand on one another and reorienting each with a random rotation in space. Flexible alignments were initiated by centering a random BCL-generated conformer of the target molecule on the native pose of the scaffold molecule and perturbing the target molecule with a random rotation in space. An alignment is considered successful if the final pose of the target molecule comes within 2.0 $\AA$ real-space symmetric RMSD of its native binding pose (Chan and Labute, 2010; Chen et al., 2006).

The CDK2 dataset was comprised of 57 unique ligands. Rigid alignment of the CDK2 system by BCL::MolAlign was comparable to results obtained via MOE (38% and 40% native pose recovery, respectively), and superior to those achieved with either ROCS or FLEXS (30% and 25%, respectively). Flexible alignments were similar across each method, ranging from 20-22%. After excluding self-aligned molecule pairs from the NxN alignment matrix, the best scoring alignment of each of the CDK2 ligands was able to recover 44 of 57 ligands less than 2.0 $\AA$ from the native binding pose (Table 6.2).

The HIV dataset contained 28 unique ligands all of which have at least ten rotatable bonds, and 16 of

Dataset	ROCS	FLEXS	MOE	BCL	ROCS	FLEXS	MOE	BCL
	Rigid				Flexible			
CDK2	30%	25%	40%	38%	20%	21%	22%	21%
HIV	39%	24%	**85%**	56%	6%	8%	16%	**22%**
P38	27%	27%	43%	46%	22%	24%	30%	31%
ESR1	44%	47%	59%	57%	25%	28%	41%	**46%**
Trypsin	57%	73%	**80%**	61%	55%	29%	61%	61%
Rhinovirus	50%	52%	50%	50%	50%	50%	50%	50%

Dataset	MOE	BCL	MOE	BCL
	Rigid		Flexible	
P38 Pharm 1	100%	100%	94%	94%
P38 Pharm 2	73%	73%	53%	45%
ESR1 Pharm 1	**94%**	86%	72%	**83%**
ESR1 Pharm 2	92%	92%	65%	**82%**

Table 6.2: **Pairwise alignment of ligands across benchmark datasets in (Labute et al., 2001; Chan and Labute, 2010)** Comparisons between four small molecule alignment methods on rigid and flexible alignment. Rigid alignment comparisons utilized the crystallographic native binding pose of each ligand as input. Flexible alignments began with a randomly generated conformer of the target molecule. In all flexible alignments the target molecule was aligned to a rigid molecule in its crystallographic native binding pose. Bolded values indicate categories in which one method recovered at least 5% of the total more native binding poses than the next best method.

which have 18 or more rotatable bonds, representing a challenging application for molecular alignment. MOE recovered 85% of the natively bound poses for the HIV ligand set via rigid alignment and 16% via flexible alignment, a considerable advancement over methods such as ROCS and FLEXS, which recovered 39% and 24% in rigid alignment, and 6% and 8% in flexible alignment, respectively. BCL::MolAlign was able to recover 55% of native poses in rigid alignment, and 22% in flexible alignment. Despite recovering fewer native poses than MOE via rigid alignment, BCL::MolAlign recovered more of the native binding poses during flexible alignment than all other methods (Table 6.2). This may be because BCL::MolAlign is able to assemble hundreds of possible conformers rapidly from a CSD-derived fragment library using BCL::Conf. Subsequent selection and refinement of these conformers with discrete bond rotations during alignment may be a more effective sampling strategy than relying on conformer sampling explicitly during the alignment stage. The bond align movers are crucial to our recovery of HIV-binding poses. We recovered only 8% of the natively bound HIV ligand poses during flexible alignment when our moves consisted of only rotation, translation, conformer swap, and bond angle perturbation. This may be because simple movers such as rotate and translate require many consecutive poorly-scoring adjustments to be made to achieve a favorable pose.

The 13 P38 kinase ligands can be divided into two pharmacophores. The first, containing the 4 ligands from PDB IDs 1M7Q, 1OUK, 1OUY, and 1OVE, is characterized by a central aromatic structure extending a piperidine/piperizine ring directly beneath the P-loop, and by a fluorinated aromatic ring accessing the back hydrophobic pocket. The second pharmacophore, represented by PDB IDs 1A9U, 1BL6, 1BL7, 1OZ1, 1W7H, 1W84, and 1YQJ, is larger with a more heterogeneous scaffold. With the exception of 1WBO, all P38 kinase ligands contain a hydrogen bond acceptor group oriented toward the backbone amide of the gatekeeper Met. In all cases with the P38 ligand set, BCL-aligned structures recovered more correct binding poses than ROCS and FLEXS. For the first pharmacophore, the BCL recovered and equivalent fraction of binding poses to MOE, with MOE achieving slightly more for the second (Table 6.2). Interestingly, the D_{max} values that give the best recovery for the P38 compounds differ from those that give the best alignments in the CDK2 and HIV datasets. This indicates that the correct D_{max} differs between datasets, and that additional optimization of D_{max} selection could further improve alignments. We also evaluated if we could improve recovery by sampling D_{max} uniformly instead of randomly. On average across the top six datasets presented in Table 6.2, uniform sampling of D_{max} between 0.70 and 1.20 recovered 1.6% fewer native binding poses, though the difference is not statistically significant.

The 13 ESR1 ligands provide another example of a single binding pocket with two distinct but overlapping pharmacophores. The first pharmacophore contains six ligands that occupy the estradiol binding-site (PDB IDs 1A52, 1GWQ, 1L2I, 1X7E, 1X7R, and 3ERD). The second pharmacophore is composed of tamoxifen-like compounds (1R5K, 1SJ0, 1UOM, 1XP1, 1XP9, 1XQC, and 2BJ4). In each of these pharmacophores,

BCL flexible alignment recovered an equivalent or higher fraction of native binding poses compared to MOE (83% and 82% vs. 72% and 65%, respectively). Of all the alignment methods, BCL was able to recover the highest fraction of native binding poses in the combined ESR1 dataset (Table 6.2).

There are seven ligands in the trypsin dataset, of which five share a near-identical binding mode, and BCL::MolAlign was able to recover their native binding poses in all of the 5x5 alignments. The remaining two ligands differ in size and binding mode, respectively. Despite these differences, during flexible alignment we achieve 61% recovery of the 7x7 matrix, on par with the recovery of MOE flexible alignment.

Finally, the rhinovirus ligand set contains eight nearly symmetric ligands with heterocyclic rings connected on either end by a long alkyl linker. As was previously discussed13, each ligand binds in two positions each of which is an inversion of the other. In this study, as in previous benchmarks, four ligands crystallized in each binding mode were used (PDB IDs 2RM2, 2RR1, 2RS1, and 2RS3 in one binding mode, and 2R04, 2R06, 2R07, and 2RS5 in the other). Successful alignment of a ligand in binding mode one to a ligand in the inverted binding mode would not be evaluated as a correct alignment using the current metric. Therefore, the maximum score for this dataset is 50%. Each alignment method including BCL::MolAlign was able to recover 50%.

6.2.5 Native binding pose recovery does not require, and is only weakly assisted by, high substructure

We investigated the extent to which maximum common substructure similarity between the target molecule and its scaffold influenced recovery of the native binding pose of the target molecule on the AstraZeneca Overlays Validation Set (1464 molecules from 121 targets). We hypothesized that the best alignments would be between molecules that shared a high degree of 2D similarity. Across all alignment pairs in the dataset, there is a weak negative correlation between native binding pose recovery and maximum common 2D substructure similarity between molecule pairs (R^2=0.17, slope = -6.34). Considering only the best alignment pair per target molecule (R^2=0.15, slope = -1.89), or only the alignment pairs where the native binding pose of the target molecule was recovered at 2.0 $\AA$ (R^2=0.13, slope = -0.67), the correlation becomes slightly weaker. These results suggest that higher 2D similarity can increase the likelihood of recovering the native binding pose, but that BCL::MolAlign recovers a large fraction of native binding poses by aligning dissimilar molecules.

We also investigated whether or BCL::MolAlign converged on energetically unfavorable conformations. For each pairwise alignment in the AstraZeneca Overlay Set benchmark, we computed the BCL::Conf score for the target molecule (i.e. the molecule being aligned to the rigid comparator). For each target molecule, we also generated conformers with BCL::Conf using the same settings that are run in the alignment protocol, and selected the single highest scoring (worst) BCL::Conf conformer. Overall, there were zero cases in

which the conformer selected from alignment had a worse BCL::Conf score than pure BCL::Conf conformer generation. We also evaluated the mean difference between the alignment conformers and either the (1) worst BCL::Conf conformer, or (2) the native conformer. The resulting mean BCL::Conf score differences are -0.24 and 0.16, respectively, suggesting that overall BCL::MolAlign conformers converge on marginally more favorable poses than those generated strictly by BCL::Conf, but that they are not always as favorable as native conformers. This latter observation is not unexpected, and overall these findings suggest that the alignment conformers represent reasonable molecule conformations.

6.2.6 BCL::MolAlign outperforms docking and substructure-based alignment in recovery of receptor-bound poses of congeneric ligands

In the later stages of drug discovery, protein-ligand docking is often employed to inform further derivatization of lead compounds. Accurate ranking of the small molecules based on their affinity depends on their accurate placement in the protein binding-pocket. Here, we compared the speed and accuracy of BCL::MolAlign to RosettaLigand on 20 unique datasets each with 4-8 congeneric ligands bound in the same protein binding pocket with a similar binding mode47. RosettaLigand is a fully flexible protein-ligand docking program distributed with the Rosetta software package, which is competitive with other state-of-the-art docking programs (Fu and Meiler, 2018; DeLuca et al., 2015; Lemmon and Meiler, 2012; Kaufmann and Meiler, 2012; Davis and Baker, 2009; Davis et al., 2009; Meiler and Baker, 2006). We employed BCL::MolAlign to align each target ligand to a scaffold ligand from each dataset. We took the geometric centroid of the same scaffold ligand as the starting position for RosettaLigand docking trials. The scaffold ligands were selected based on chronology of earliest deposition in the Protein Data Bank (PDB). All alignment and docking trials were performed starting from randomly generated ligand conformers. In this way, the benchmark emulates a realistic drug discovery process, in which the binding mode of the single earliest co-crystalized complex guides virtual screening.

Across all datasets, the top-scoring RosettaLigand model by protein-ligand interaction score for each protein-ligand complex was within 2.0 $\AA$ of the experimentally determined binding poses in 60% of cases. In contrast, the top-scoring model by property distance to the scaffold ligand in BCL::MolAlign identified the correct binding pose in 86% of cases (82% of cases when self-alignments are excluded). To generate one model with RosettaLigand using the protocol described in the Methods section takes approximately 90 seconds. A typical docking run requires approximately 102 - 103 independent docking trials (Fu and Meiler, 2018; DeLuca et al., 2015) to produce a native-like binding pose. In this benchmark, we generated 1000 models for each dataset. In comparison, a single alignment with five serial independent trajectories in BCL::MolAlign takes on average approximately 46 seconds (9.2 seconds per trajectory, single CPU thread)

on Intel Xeon X5690 processors. On a 12-core workstation, for example, this allows screening of approximately 45,000 ligands against a single scaffold ligand in 24 hours.

Performance on 3 of the 20 datasets in particular (HCV, TPPHO, and CTAP) was previously found to be improved by simultaneous docking of the ligands within each binding pocket compared to traditional docking47. We found that BCL::MolAlign similarly provides an advantage over RosettaLigand docking in these datasets. This is most clear in the HCV dataset (Figure 6.5, row 1). The binding pocket is large with multiple favorably scoring binding modes, and in only 2/6 cases did RosettaLigand recover a native-like binding pose as the top-scoring model. In contrast, BCL::MolAlign was able to recover native-like poses in 5/6 cases by superimposing to the earliest available scaffold (PDB ID 3BR9). Similarly, failure of RosettaLigand to properly place the core bi-substituted aromatic ring structure occurred systemically in the CTAP dataset here and elsewhere (Fu and Meiler, 2018) (Figure 6.5, row 3). The resultant translational error caused RosettaLigand to only recover native-like binding poses in 3/6 cases, while BCL::MolAlign accurately recovered 6/6.

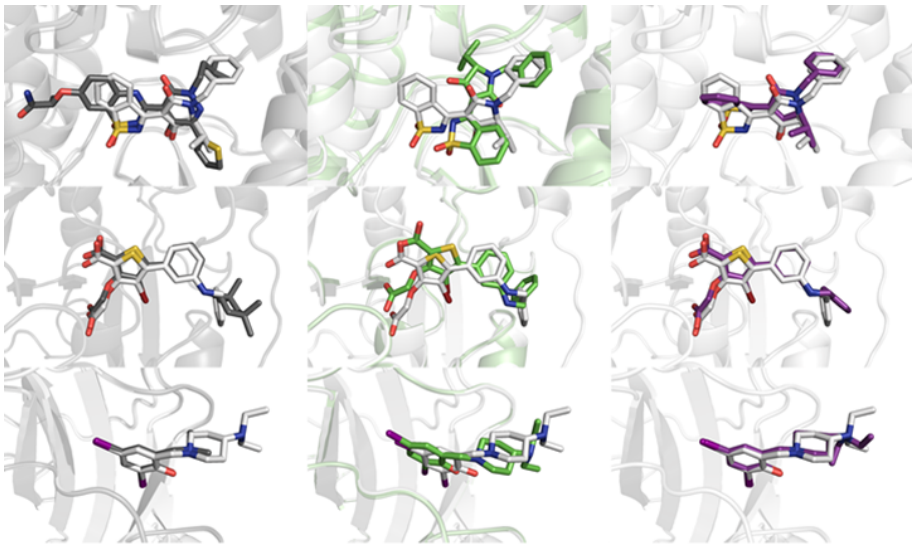


Figure 6.4: Visual representations of docked versus aligned poses in challenging docking targets. Comparisons show the protein-ligand complexes of the crystallized scaffold (gray) and crystallized target (white) molecules (A). The crystallized pose of the target molecule (white) is also shown with the RosettaLigand docked pose (green; B) and the BCL::MOLALIGN flexibly aligned pose (purple; C). Examples correspond to molecules from the HCV (row one), TPPHO (row two), and CTAP (row three) datasets.

Given the high degree of substructure similarity between the ligands in each congeneric set, an important question is whether or not BCL::MolAlign provides a benefit over a substructure-based alignment method. To test this, we generated 100 conformers of each ligand with BCL::Conf and aligned all conformers to their respective scaffold molecules based on maximum common substructure. Substructure-based alignments were performed with BCL alignment tool AlignToScaffold (ATS) as described in Methods. First, we compared

the abilities of BCL::MolAlign and ATS to recover native binding poses when the input target molecule was the native conformation. Across all 20 datasets, BCL::MolAlign recovered 96% of the native binding poses, while ATS recovered 89% (Table 6.3).

		Maximum Common Substructure Alignment				BCL::MolAlign	
Dataset	Total	Native Conformer	RMSD	ChargeRMSD	MolAlign Score	Native Conformer	Flexible Alignment
AR	5	5	5	5	5	5	5
AVGLU	5	5	5	4	4	5	4
BETAX	5	5	5	5	5	5	5
CALM	5	4	5	5	5	5	5
CATB	6	6	5	5	5	4	6
CDK2	7	6	5	7	3	7	3
CTAP	6	6	3	3	3	6	6
FXA	4	4	3	4	3	3	3
GLCB	5	2	2	2	4	5	5
HCV	6	4	4	4	6	6	6
HSP90	6	3	2	2	5	6	6
LPXC	5	5	5	5	5	5	5
MTAN	7	5	2	3	5	7	5
P38	4	4	2	3	2	4	3
PNMT	5	5	2	2	2	5	2
RET	4	4	4	4	4	4	4
SYK	8	8	6	6	6	8	7
THERM	7	7	6	6	5	6	5
THROM	5	5	5	5	4	5	5
TPPHO	6	6	3	3	6	6	5
Total	111	99	79	83	87	107	95
% Recovery		89%	71%	75%	78%	96%	86%

Table 6.3: **Comparison between BCL::MolAlign and maximum common substructure-based alignment of congeneric ligands.**

Next, we utilized multiple scoring metrics to try and optimize recovery of the native pose with ATS. To evaluate which conformer of the target ligand yielded the best fit to the scaffold, we used an RMSD100-like metric (Gregory et al., 2014). With this scoring system, we were able to recover 71% of the native binding poses. Subsequently, we used a property-weighted version of the RMSD100-like metric (previously termed "ChargeRMSD") (Gregory et al., 2014), and improved recovery of the ATS alignments to 75%. Finally, we performed ATS and scored the resultant alignments with the BCL::MolAlign scoring system, with which we again improved recovery to 78%, but was still below the 86% recovery of BCL::MolAlign (Table 6.3).

6.2.7 Discussion

In summary, we have developed a novel small molecule flexible alignment algorithm called BCL::MolAlign. BCL::MolAlign utilizes multi-tiered MCM sampling to superimpose and flexibly refine molecular conformers according to a customizable property-based metric. It combines established molecular conformer generator capabilities with on-the-fly dihedral angle optimization for refinement. We have benchmarked BCL::MolAlign against state-of-the-art commercial and free software. Generally, BCL::MolAlign performs on par with, or superior to, similar software packages. Alignments generated with BCL::MolAlign can serve as pharmacophore hypotheses, aid in the selection of ligand conformers and starting poses for protein-ligand docking, and identify likely 3D conformers based on template compounds. When a starting binding pose is available for a protein-ligand complex, BCL::MolAlign is capable of identifying native-like binding poses for

large libraries of small molecules in parallel. We demonstrate how BCL::MolAlign can be used to improve the efficacy of ensemble docking programs including RosettaLigandEnsemble (Fu and Meiler, 2018). Moreover, we have demonstrated that the BCL::MolAlign alignment score has predictive value and can be used to distinguish active from inactive compounds. As an extension to this finding, we also anticipate that the alignment score could make a valuable descriptor in QSAR models. Finally, BCL::MolAlign was designed to facilitate high-throughput screening of small molecule libraries. It is "embarrassingly parallel" in its implementation, allowing independent alignments to occur simultaneously across multiple threads. As a result, BCL::MolAlign is fit for medium- to high-throughput application projects in academia and industry.

6.3 Methods

6.3.1 Benchmarking Dataset Preparation

The CDK2, HIV, P38, ESR1, trypsin, and rhinovirus datasets comparisons were assembled from the PDB IDs in (Chan and Labute, 2010). Protein-ligand co-crystal structures were superimposed by Cα atoms of the ligand binding pocket in PyMOL (DeLano, 2007). The positions of each ligand in the protein-ligand co-crystal structure alignment were taken to be native/scaffold poses. The DUD datasets prepared for the virtual screening comparisons were obtained from the Kihara Lab at http://kiharalab.org/ps_ligandset/ . The cognate ligands provided with each dataset were taken to be the scaffolds. All datasets used in the benchmark are available in the Supplementary Material. Comparisons with RosettaLigand were completed using 20 protein-ligand datasets from Fu Meiler (Fu and Meiler, 2018). Target ligands were assigned a random 3D conformer prior to flexible alignment to scaffolds. Tanimoto largest common substructure comparisons were performed in the BCL. Substructures were defined by maching atoms by atomic numbers and bonds by bond order (with aromatic bonds given a distinct bond order), and ring membership.

6.3.2 Chemical Properties

All BCL::MolAlign alignments were performed with the same set of chemical properties. For each atom type we computed Gasteiger partial charges (Gasteiger and Marsili, 1980), polarizability (Miller, 1990), electronegativity (Pauling, 1932), hydrophobicity (Labute, 2000), Van der Waals volume (RN2, a), aromaticity (RN2, b), hydrogen bond donor (OH/NH), and hydrogen bond acceptor (O/N) status. Aromaticity is calculated as the Marvin General method (RN2, a), which has similarities to the more common Daylight method (RN2, b). Electronegativity values are determined by element type from standard periodic table values. As in Chan Labute 2010, those atoms which are at least two bonds away from a hydrogen-bonding atom are designated as hydrophobic (Chan and Labute, 2010). Property weights were obtained as previously described by computing the inverse standard deviation of each property's occurrence across a sample library of drug-like

small molecules (Gregory et al., 2014; Butkiewicz et al., 2013).

6.3.3 Alignment Parameters

BCL::MolAlign is based on a Monte Carlo – metropolis architecture. Accordingly, random moves are scored and accepted if they either (1) improve upon the existing score, or (2) fail to improve the existing score but win a "coin toss" with a probability of winning that is dependent on the change in score and on the temperature of the simulation38. Higher temperatures increase the likelihood of a move being accepted. We utilize a temperature-control system which automatically adjusts every 10 iterations such that the initial acceptance rate at the beginning of the simulation is 50% and the final acceptance rate is 1%. The target ratio adjusts linearly over the course of a trajectory.

First, all molecules are assigned explicit hydrogen atoms and Gasteiger atom types55. Next, a random 3D molecular conformer is generated for each molecule with BCL::Conf (Kothiwale et al., 2015). BCL::MolAlign alignments were performed with the following parameters: 100 conformers were generated for each molecule except for those in the CDK2 and HIV datasets for which 500 and 2000 conformers were generated, respectively; the number of conformer pairs is set equal to the number of conformers for the purposes of this benchmark; 400 iterations were performed for the MC Optimization Tier 1 but terminated early if the score failed to improve after 160 consecutive iterations; 600 iterations were performed for the MC Optimization Tier 2 but terminated early if the score failed to improve after 240 consecutive iterations; 200 iterations were performed for the MC Optimization Tier 3 but terminated early if the score failed to improve after 80 consecutive iterations; 5 independent trajectories with random maximum atom distances between 0.70 and 1.20 $\AA$; re-scoring to normalized maximum atom distances was completed on the top 5 molecules from each independent trajectory; a mismatch penalty constant of 2.0 was used throughout. Collectively, these values are specified as the default settings in BCL::MolAlign, with the exception of the number of conformers and conformer pairs, which have been set to default values of 500 and 100, respectively. These settings are also implemented as defaults in the BCL::MolAlign webserver. For additional details and command-lines, see Supplementary Methods. Performance benchmarks comparing RosettaLigand with BCL::MolAlign were completed on Intel Xeon X5690 processors using a single CPU thread per process.

CHAPTER 7

General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps

This chapter is taken from Brown, B. P.; Mendenhall, J.; Geanes, A. R.; Meiler, J. J. Chem. Inf. Model. 2021, 61 (2), 603–62017.

7.1 Introduction

Computer-aided drug discovery (CADD) is a broad category of methods that can be employed to increase the efficiency of the drug discovery process. Broadly, CADD methods can be subdivided into two categories: ligand-based (LB) and structure-based (SB) (Sliwoski et al., 2014). LB methods predominantly employ similarity metrics to compare ligands with known biological activity or chemical attributes to a library of prospective small molecules. Among the most widely used LB methods are quantitative structure-activity relationship (QSAR) models, which relate quantitative chemical descriptors of molecules to known biological activities (Sliwoski et al., 2014; Leelananda and Lindert, 2016). QSAR models lend themselves to supervised machine learning methods, such as artificial neural networks (ANN) and Random Forest (RF) (Butkiewicz et al., 2013; Dahl, 2014; Mendenhall and Meiler, 2016; Hillebrecht and Klebe, 2008; Manchester and Czerminski, 2008; Svetnik et al., 2003). Indeed, over the last two decades we have demonstrated the efficacy of ANNs in LB classification tasks compared to other methods, such as support vector machines, and employed them to identify multiple G-protein-coupled receptor (GPCR) allosteric modulators (Butkiewicz et al., 2013; Geanes et al., 2016; Lowe et al., 2010; Mueller et al., 2010; Bleckmann and Meiler, 2003). In that time, we have contributed to multiple aspects of QSAR method development, including early efforts to expedite model training with GPU programming (Lowe et al.), chemical descriptor and toolkit development (Sliwoski et al., 2012, 2015; Mendenhall et al., 2021), improving QSAR ANN architectures with dropout (Mendenhall and Meiler, 2016), and dataset assembly for community benchmarking (Butkiewicz et al., 2013). We have accomplished this largely with the development of the BioChemical Library (BCL), a primarily ligand-based academic open-source cheminformatics toolkit (Brown et al., 2022). LB methods can often rank compounds many orders of magnitude faster than SB methods. Despite being very rapid and easily deployed on large databases for virtual high-throughput screening (vHTS), ligand-based methods have inherent limitations. Most notably, LB methods make predictions in the absence of binding pocket information. As a result, predictions made from LB methods must be target-specific, and generating LB models for a given target, especially QSAR models, may require a large amount of model training data. Thus, there is considerable

interest in developing target agnostic, rapid SB methods for vHTS.

SB methods provide information about small molecule interactions with the binding pocket. Critically, this should allow SB-methods to be target agnostic and provide chemically meaningful insight with which to guide hit optimization. Unfortunately, the most accurate SB methods come with a computational cost prohibitive for vHTS. Accurate prediction of small molecule binding affinities to target proteins is a key challenge in SB-CADD. Structure-based alchemical free energy approaches, such as free energy perturbation (FEP) and thermodynamic integration (TI), are widely considered to be the most accurate (Wang et al., 2019a; **?**; Zou et al., 2019). Other approaches, such as molecular mechanics Poisson-Boltzmann or Generalized-Born surface area (MM/PB(GB)SA), or protein-ligand docking semi-empirical scoring functions, can also provide reliable relative binding free energies, but with overall performance seemingly being more system-dependent (Tokudome et al., 2020; Wang et al., 2019c; Sun et al., 2018). Faster, but less accurate, docking score functions are being increasingly scaled to medium- and high-throughput virtual screening (Stein et al., 2020; DeLuca et al., 2015).

In the last decade, many machine learning approaches have been developed to increase the speed and accuracy of SB virtual screening approaches. As early as 2010, random forest (RF) rescoring of docked poses demonstrated that machine learning algorithms could provide rapid and competitive prediction of protein-ligand binding affinities (RF-Score) (Ballester and Mitchell, 2010). A variation on RF as a modeling tool for protein-ligand binding affinity prediction is ΔVinaRF, which uses random forest (RF) to predict an error correction term for the AutoDock Vina docking score function (Wang and Zhang, 2017). More recently, deep learning with convolutional neural networks (CNN) has been widely investigated to predict binding affinities. For example, DeepVS is a CNN that attempts to generalize binding mode information by encoding local atomic neighborhoods around each selected ligand atom using simple descriptors (i.e. atom types, charges, distances, and interacting amino acid identity) (Pereira et al., 2016). Multiple grid-based CNNs have also been developed, such as KDEEP and a CNN by which Ragoza et al., which treat protein-ligand complexes as 3D images colored by specific atom type and pharmacophore properties (Ragoza et al., 2017; Jiménez et al., 2018). AtomNet is another grid-based CNN that also includes features derived from protein-ligand interaction fingerprints (Izhar Wallach, 2015).

It is well-known that cheminformatics machine learning algorithms can be strongly limited in their domain of applicability by the chosen training set and descriptors (Minovski et al., 2013; Sheridan, 2012; Tetko et al., 2008; Schroeter et al., 2007; Ruiz and Gómez-Nieto, 2018; Roy et al., 2015; Carrió et al., 2014). There is concern that some newer CNN techniques demonstrating exceptional performance may suffer from lack of generalizability owing to dataset and training biases (Ragoza et al., 2017; Sieg et al., 2019). Even in cases where machine learning models make accurate predictions, the chemical basis of these predictions is not

easily interpreted without substantial input sensitivity and feature analysis. This infamously gives rise to the "black box" problem of machine learning algorithms, especially deep neural networks (DNNs).

Finally, a major motivation for the current project is to incorporate a modular and customizable SB score function into the BCL for use in the ongoing development SB design algorithms. Currently, the BCL is only able to support LB design algorithms. Ultimately, we anticipate that increasing the capabilities of the BCL to perform both LB and SB design tasks will make it a valuable companion to other academic molecular modeling software projects, such as the Rosetta macromolecular modeling and design software suite (Leman et al., 2020).

To address these issues, we have designed a novel SB protein-ligand binding affinity and pose prediction model based on distance-dependent signed atom property protein-ligand correlations. Instead of encoding specific protein and ligand properties, our method encodes the protein-ligand interaction feature space. This is analogous to the formation of statistical pair potentials, except that here we do not formally provide any constraints on the function to be approximated. We demonstrate that fully-connected feed-forward neural networks trained with our new descriptors are competitive with existing state-of-the-art machine learning methods and docking methods at protein-ligand binding affinity prediction, pose prediction, and virtual screening power. Moreover, we explicitly demonstrate that the performance of our models is not dependent on exploiting dataset bias. Finally, we show how our models can be rapidly decomposed into human interpretable pharmacophore maps. These pharmacophore maps allow users to visualize the atoms/substructures of their molecules that drive the activity prediction, as well as map predicted or known relative binding free energy changes across molecule ensembles to specific substructures. This will be the first SB scoring tool available in the BCL, and the pharmacophore mapping tool is fully compatible with the LB QSAR methods currently implemented. Together, we believe these tools improve the utility of the BCL for SB hit identification and lead optimization in drug discovery.

The new descriptors, models, and pharmacophore mapping application will be available in the upcoming BCL version 4.1 release, an academic open source software package for cheminformatics written in the C++ programming language. It is our hope that our new method will be used in conjunction with other advancements in machine learning-based QSAR/QSPR to continue to improve the efficiency of drug discovery.

7.2 Results

7.2.1 On the development of a pose-dependent protein-ligand property correlation descriptor

Currently, the top-performing deep learning scoring algorithms that predict binding affinities from protein-ligand complexes are CNNs that encode neighboring ligand and receptor atoms spatially and/or chemically (e.g. hydrogen bond donor/acceptor heuristics) (Ragoza et al., 2017; Jiménez et al., 2018). One critique

of these CNNs is that test-set performance can be attributed to learning ligand-specific features and not the protein-ligand interface features (Sieg et al., 2019). In other words, the neural network can perform well on the tests simply by learning the biases in the ligand datasets. To avoid any such potential limitations here, we developed a pose-dependent protein-ligand interaction descriptor based on sign-aware 3DAs. This descriptor can be likened to a potential of mean force profile in which the collective variables are the pairwise interatomic distances between protein and ligand atoms for specific chemical properties/heuristics.

7.2.2 Small molecule chemical property autocorrelations

Consider a property-weighted 3D autocorrelation (3DA) function for a single small molecule. An atom-based property allows the 3DA to represent the spatial distribution of properties of interest:

$$3DA(r_a, r_b) = \sum_{j}^{N} \sum_{i}^{N} \delta\left(r_a \leq r_{i,j} < r_b\right) P_i P_j e^{-\beta {r_{i,j}}^2} \tag{7.1}$$

where r_a and r_b are the boundaries of the current distance interval, N is the total number of atoms in the molecule, r_(i,j) is the distance between the two atoms being considered, δ is the Kronecker delta, β is a smoothing parameter referred to as 'temperature' (Sliwoski et al., 2012; Hemmer et al., 1999), and P is the property computed for each atom. 3DAs computed for signed properties (e.g. partial charge) contain, for each distance interval, three values corresponding to product sums of each of the three possible sign pairings (-/-, +/+, -/+) (Sliwoski et al., 2015).

7.2.3 Recasting property space into protein-ligand interaction distance bins

Instead of corresponding to intramolecular atomic distances, the distance bins now correspond to intermolecular protein-ligand interatomic distances. The property correlation is between each atom in the ligand and all atoms in the receptor within a specified radius (Figure 7.1):

$$PLC(r_a, r_b) = \sum_{l}^{N_{lig}} \sum_{p}^{N_{prot}} \delta\left(r_a \leq r_{l,p} < r_b\right) P_l P_p e^{-\beta_i} \tag{7.2}$$

where r_a and r_b are the boundaries of the current protein-ligand interatomic distance interval, N_lig and N_prot are the total number of atoms in the ligand and receptor, respectively, r_(l,p) is the distance between the current protein-ligand atom pair, δ is the Kronecker delta, β is the temperature, and P_l and P_p are the properties computed for ligand and receptor atoms l and p, respectively. As with 3DA in (eq. 1), PLC (protein-

ligand correlation) descriptors distinguishes signed pairs, but can also optionally include an additional bin (–/++/-+/+-) to account for opposite sign pairings between the protein and the ligand (Figure 7.1A). This can be useful if the properties between which the correlations are being taken are not identical, or if the model being built is leveraging pre-existing knowledge about the chemical makeup of the system in study.

For example, consider the descriptor "HBondDonorTernary". This descriptor returns a 1 if an atom is a hydrogen bond donor, -1 if it is a hydrogen bond acceptor, and 0 otherwise. One could choose to differentiate hydrogen bond donor/acceptor pairs between the protein and the ligand (e.g. asymmetric: -+/+-), or to group all opposite sign pairs together (symmetric -/+). Sign pair discrimination is illustrated in Figure 7.1A for a property that tracks the protein-ligand directionality of opposite sign pairings. We empirically chose a total distance of 7.0 $\AA$ discretized at 0.50 $\AA$ intervals, resulting in either 42 (symmetric) or 56 (asymmetric) values per property (see subsection on feature parameterization in Methods and Supporting Information).

7.2.4 Representing protein-ligand interactions with property correlation descriptors

PLC descriptors (eq. 2) encode interactions between protein-ligand atomic atoms as represented by a variety of atomic properties: partial charge, electronegativity, polarizability, hydrophobicity, hydrogen bond donors and acceptors, aromatic and generic ring membership, heavy and light atoms. These atomic features are a superset of those we used previously for QSAR (Sliwoski et al., 2015; Mendenhall and Meiler, 2016), and are identical to those we used previously for superimposition of similar molecules (Brown et al., 2019b).

To mitigate feature redundancy, we summed feature interactions that were nominally equivalent. For example, consider the PLC descriptor that represents the signed correlation between atomic partial charges in receptor and ligand atoms: 3DAPairRS050(Atom_SigmaCharge). In this descriptor, we summed -/+ (ligand negative charge, protein positive charge) with +/- (ligand positive charge, protein negative charge) interactions under the notion that these are equivalently favorable pairings. We took a similar approach for hydrogen bond donation, hydrophobic interactions, and heavy atom / hydrogen atom discrimination. Some descriptors, such as polarizability and electronegativity, are strictly positive valued, and therefore do not require binning by sign pairs.

While each of the previously mentioned descriptors can be considered symmetric in that we are correlating the same property for both the receptor and the ligand (e.g. partial charge), interactions can also be described by complementary interactions between dissimilar chemical properties. For example, interactions between aromatic ring systems and polar vs. hydrophobic atoms. To create a property that can describe this interaction, we need to utilize Atom_HydrophobicTernary, which is an atom property that encodes hydrophobic atoms as +1, and polar atoms as -1. To better distinguish highly polar from less polar atoms, we multiply Atom_HydrophobicTernary by polarizability. We then encode aromatic-polar, aromatic-hydrophobic interac-

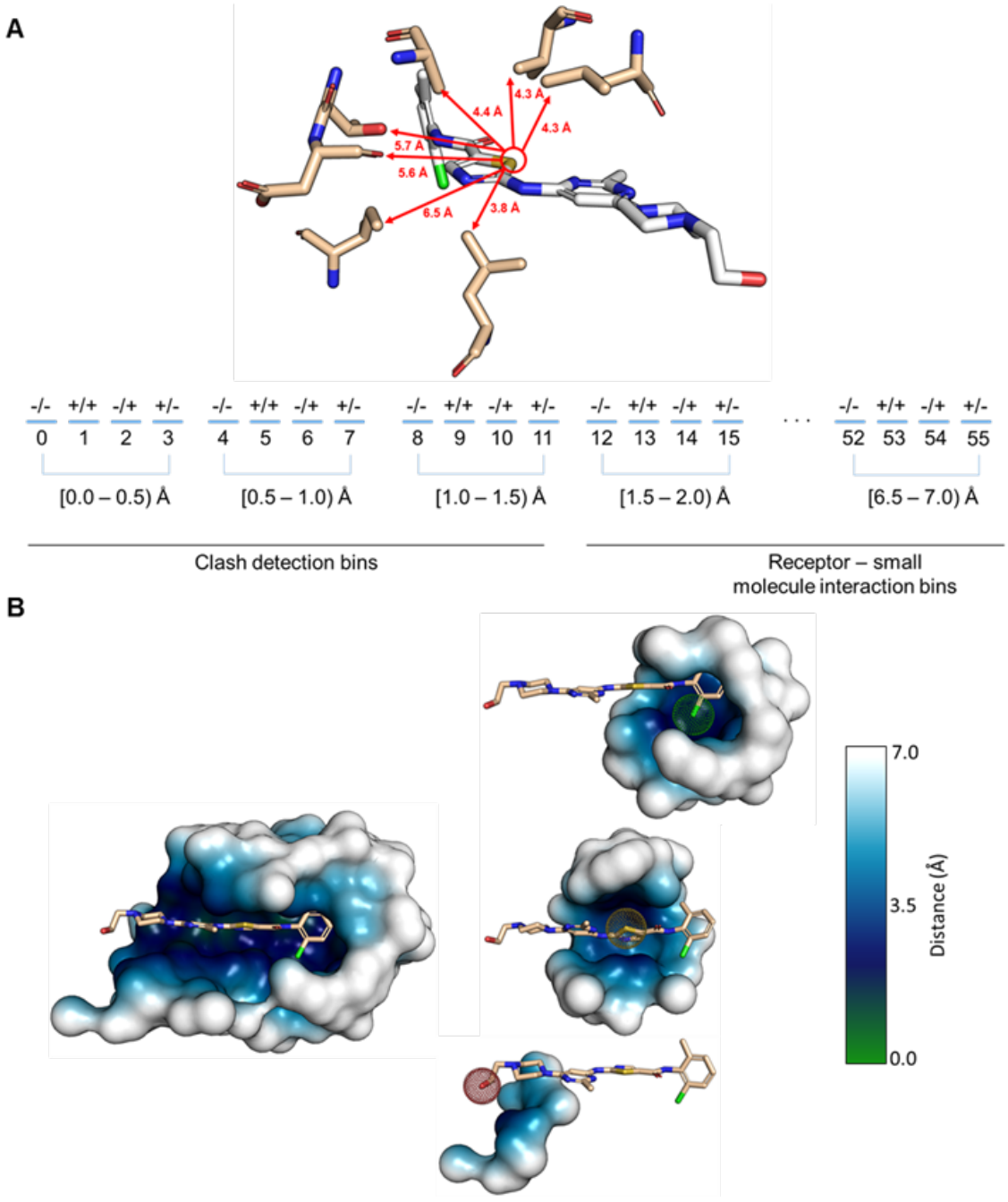


Figure 7.1: Schematic of pose-dependent protein-ligand descriptor. (A) Schematic representation of pose-dependent protein-ligand interaction feature space. (B) Surface representation of discoidin domain receptor 1 (DDR1) kinase binding pocket heavy atoms within 7.0 $\AA$ of select atoms within dasatinib. The surface representation is colored by distance to the selected atom. Dasatinib shown in stick configuration colored by element type with the selected atom indicated by dot sphere.

tions with the PLC descriptor, "3DAPairRSAsym050(Multiply(Atom_HydrophobicTernary, Atom_Polarizability), Atom_IsInAromaticRingTernary)". In this descriptor, each distance bin is further discretized into -/- (ligand polar atom polarizability with a non-aromatic receptor atom), +/+ (ligand hydrophobic atom polarizability with an aromatic receptor atom), -/+ (ligand polar atom polarizability with an aromatic receptor atom), and +/- (ligand hydrophobic atom polarizability with a non-aromatic receptor atom). An inverted version of this descriptor, in which hydrophobicity is with respect to the receptor and aromaticity to the ligand, is also employed here.

With these features, we trained two neural networks. BCL-AffinityNet is a "deep" single-task neural network (2 hidden layers, 512 neurons in the first hidden layer and 32 neurons in the second layer) to directly predict log-scaled protein-ligand binding affinity values. BCL-DockANNScore is a multi-tasking shallow neural network (1 hidden layer with 32 neurons) that classifies binding poses as less-or-equal to 1.0, 2.0, 3.0, 5.0, or 8.0 $\AA$ from the native (co-crystallized) binding mode. Both of these models utilize only PLC descriptors (eq. 2), with BCL-DockANNScore including an additional PLC descriptor that discretizes hydrogen bond donor/receiver pair angles.

Finally, we note that we did not perform a deep exploration of possible base chemical descriptors and there are likely many additional features that could be effective (e.g. explicit consideration of π-interactions, σ-hole interactions, transition metal properties, solvation energies, etc.). Additionally, we did not perform feature selection to optimize the performance of our model on the benchmark training sets to avoid potentially over-optimizing the models for the training data. For a detailed evaluation of the importance of each feature in BCL-AffinityNet and BCL-DockANNScore, please see the top 20 features by model input sensitivity and a decomposition of each descriptor into the average input sensitivity per sign pair (Figure S1 – S8) in the Supporting Information.

7.2.5 Scoring power evaluation of BCL-AffinityNet

We trained BCL-AffinityNet on protein-ligand complexes from the PDBbind v.2016 refined set and all general set protein-ligand (small molecule) complexes for which binding constants were available. Protein-ligand pairs comprising the coreset (285 unique test set complexes) were entirely excluded from training. BCL-AffinityNet was trained with descriptors of the form (eq. 2). See the Supporting Information for a sample feature code object file and command-lines to generate the model.

We first tested the performance of BCL-AffinityNet on the scoring power task described in the comparative assessment of score functions 2016 update (CASF2016). This task evaluates affinity prediction across the PDBbind v.2016 coreset comprised of 285 protein-ligand pairs on 57 targets (5 small molecules per target) by measuring the Pearson correlation coefficient (R) between predicted and experimental values. It has

previously been noted that binding affinities in this task correlate strongly with both the fraction of buried solvent accessible surface area (ΔSAS, R=0.63) (Figure 7.2A)1 and several scalar ligand descriptors, including molecular weight (MW, R=0.50), topological polar surface area (TPSA, R=0.20), logP (R=0.32), and polarizability (R=0.52). An important measure of success is whether or not the affinity prediction method is capable of performing better than these simple metrics that are unaware of specific protein-ligand interactions.

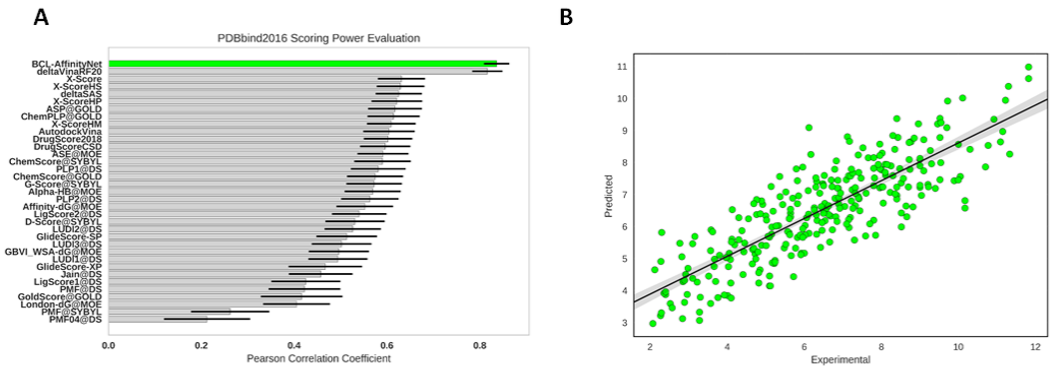


Figure 7.2: Scoring power evaluation of BCL-AffinityNet. (A) Comparison of BCL-AffinityNet scoring power to other methods from the CASF2016 benchmark by Su et al.1. Error bars indicate the 90% confidence interval (B) Linear regression of experimental vs. predicted pKd values in the CASF2016 coreset.

BCL-AffinityNet is among the best algorithms on the scoring power task (R=0.84) (Figure 7.2A, B). ΔVinaRF¬20, which is a protein-ligand interaction score function that uses a random forest (RF) algorithm to predict an error correction term on the AutoDock Vina score, performed similarly on the original CASF2016 report (Figure 7.2A)1. However, as reported previously1, the training set for ΔVinaRF includes 140 of the coreset test complexes. Lu and colleagues re-evaluated the scoring power of ΔVinaRF after retraining it without any of the coreset complexes and found that is still performed better than ΔSAS but with worse scoring power than originally reported (R=0.73) (Lu et al., 2019).

BCL-AffinityNet performs competitively with other machine learning models, such as the grid-based CNN KDEEP (R=0.82) and RF-Score (R=0.80). We note that KDEEP was evaluated on the 290 molecule version of the PDBbind coreset, not the canonical 285 molecule set. Moreover, in the absence of the underlying distributions it is unclear if these results are statistically different; however, the effect sizes are similar.

7.2.6 Explicit assessment of dataset bias on BCL-AffinityNet scoring power performance

It is increasingly well-documented that strong machine learning model performance on QSAR tasks can be the result of dataset bias (Sieg et al., 2019; Yang et al., 2020; Chen et al., 2019). Indeed, Yang et al. found that atomic CNNs (ACNNs) trained solely on ligand or receptor pocket features performed just as well as ACNNs trained on protein-ligand complexes (Yang et al., 2020), suggesting that the model was unable to leverage

features relating to the protein-ligand interactions in a meaningful way. Therefore, we sought to determine the extent to which dataset biases may be inflating BCL-AffinityNet performance.

First, we trained a BCL-AffinityNet Y-scramble model, in which the result labels were shuffled between training examples. The Y-scramble model is a negative control, and as expected we find virtually no correlation between predicted and experimental results on the coreset with this model (Figure S9).

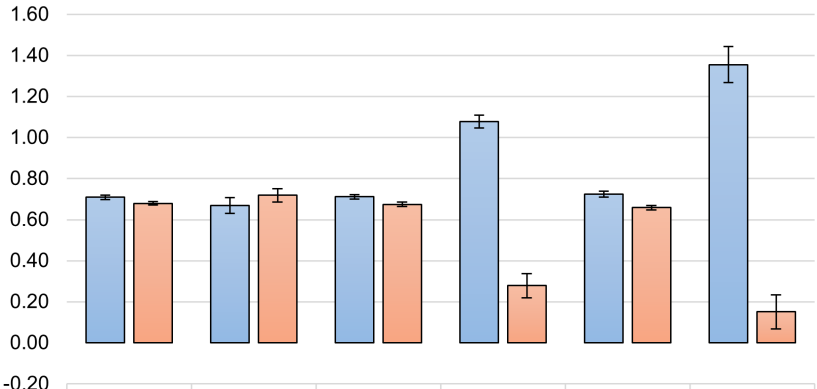
Next, we generated LB and pocket-based QSAR models with the same architecture as BCL-AffinityNet. These models were trained with the 3DA descriptor equivalent of the PLC features. In an ideal dataset, ligand and protein pocket controls would have near zero correlation to experimental results; however, consistent with the findings of Yang et al. (Yang et al., 2020), the LB and pocket-based QSAR models each had correlation coefficients greater than 0.50 at 0.72 and 0.61, respectively (Figure S9).

To assess the impact of dataset bias on our PLC models performance for out-of-class predictions, we generated three new leave-class-out test-set splits based on ligand, protein pocket, or combined ligand and protein pocket similarity to the PDBbind v.2016 coreset. Specifically, we generated a K-means (k=75) applicability domain (AD) model from the 3DAs of the ligands, protein pockets, or combination of ligands and protein pockets of the PDBbind v.2016 coreset. Using each of these AD models, we removed training samples that were further from their nearest Kohonen map node than the furthest point of the PDBbind v.2016 coreset was from the AD model. Intuitively, the new test-sets thus include only points that are outside the nominal descriptor space given by the PDBbind v.2016 coreset for ligands, protein pockets, or combination ligand-protein pockets. This has the effect of making the training set feature space more representative of the PDBbind v.2016 coreset feature space, while simultaneously creating new test sets that are outside of PDBbind v.2016 coreset feature space.

This resulted in the creation of a LB AD test set (n=995), pocket AD test set (n=379), and combined AD test set (n=1377) (see Methods for additional details). We hypothesized that the LB QSAR model would perform poorly on the LB AD test set, that the pocket-based QSAR model would perform poorly on the pocket AD test, and that both models would perform poorly on the combined AD test set. We further hypothesized that if models trained on PLC descriptors are truly generalizable SB score functions, then their performance on all three test splits ought not to be significantly worse than their training random-split cross-validation metrics.

We found that the LB QSAR models performed worse on the LB AD test set (R=0.28) than on the random-split training cross-validation sets (R=0.67) (Figure S10). Similarly, the pocket-based QSAR model performed worse on the pocket AD test set (R=0.33) than on the training splits (R=0.63) (Figure S11). We also note a reduction in performance of the pocket-based QSAR model on the LB AD test set relative to training (R=0.51 vs. R=0.64, respectively), as well as a reduction in performance of the LB QSAR model

on the pocket AD test set relative to training (R=0.54 vs. R=0.65, respectively) (Figures S10 – S11). On the combined AD test set, we observe the worst performance of the LB (R=0.28) and pocket-based (R=0.15) QSAR models (Figure 7.3).



	BCL::Affinity Net Training (5xCV)	BCL::Affinity Net Combined AD Test (n=1377)	Ligand-based DNN Training (5xCV)	Ligand-based DNN Combined AD Test (n=1377)	Pocket-based DNN Training (5xCV)	Pocket-based DNN Combined AD Test (n=1377)
Normalized Mean Absolute Error	0.71	0.67	0.71	1.08	0.72	1.36
Correlation Coefficient	0.68	0.72	0.67	0.28	0.66	0.15

Figure 7.3: Performance evaluation on the combined AD test set. A total of 1377 training samples were excluded from the initial training set of 7568 samples (see Methods for details). The remaining 6191 training samples were used to train BCL-AffinityNet (i.e. a single-task regression DNN with PLC features), a signed 3DA LB QSAR model, or a signed 3DA pocket-based QSAR model. Training was completed with five-fold random-split cross-validation. Columns and error bars represent the mean and standard deviation of NMAE (blue) or Pearson correlation coefficient (red) across either the five-fold random-split cross-validations (training) or five-fold random splits of the combined AD test set (testing).

In contrast, we observe that BCL-AffinityNet, when retrained to exclude each AD test set, consistently performs well (R=0.72, 0.75, and 0.72 for the LB, pocket-based, and combined AD test sets respectively) despite the reduced training set size and coverage (Figure 7.3, Figures S10 – S11).

To evaluate whether PLC descriptors are effective with other machine learning model types, we have utilized WEKA (Witten et al., 2016) to train a random forest version of BCL-AffinityNet (termed AffinityRF for ease) for evaluation on the PDBbind v.2016 coreset and the combined AD test split. AffinityRF achieves good correlation (R=0.79 and 0.70, respectively) on both tasks, suggesting PLC descriptors may be suitable in multiple machine learning paradigms (Figure S12). Altogether, these results suggest that the PLC descriptors encode generalized representations of protein-ligand interactions.

7.2.7 Performance evaluation on subsets of the CSAR-NRC HiQ test sets

As additional independent tests, we evaluated the performance of our models on the CSAR NRC HiQ test sets. For the purposes of a head-to-head comparison with two of the leading machine learning methods in the

field, KDEEP and RF-Score, we first compared our model to the 55 and 49 compounds of the CSAR NRC HiQ test set 1 and 2, respectively, which were previously evaluated for KDEEP and RF-Score in Jimenez et al. 32. For this evaluation, we re-trained our models with the PDBbind set as described previously, but we also excluded any of the 55 or 49 compounds found in the CSAR test set from training.

RF-Score performed the best on set 1 (R=0.78, RMSE=1.99) with KDEEP (R=0.72, RMSE=2.09) and BCL-AffinityNet (R=0.72, ρ=0.77, RMSE=2.02) performing similarly one another (Table 7.1). In contrast, BCL-AffinityNet is the top performing model (R=0.85, ρ=0.82, RMSE=1.37) on set 2, followed by RF-Score (R=0.78, RMSE=1.66) and KDEEP (R=0.65, RMSE=1.92) (Table 7.1).

		Benchmark Test Set					
		CSAR NRC-HiQ set 1 (n=55)			CSAR NRC-HiQ set 2 (n=49)		
Descriptor Set	**Model Type**	Pearson R	Spearman ρ	RMSE	Pearson R	Spearman ρ	RMSE
BCL-AffinityNet	DNN 2x512-32	0.72	0.77	2.02	0.85	0.82	1.37
Molecular Weight	Scalar Property	0.29	0.32	N/A	0.45	0.32	N/A
TPSA	Scalar Property	0.03	0.12	N/A	-0.10	-0.10	N/A
LogP	Scalar Property	0.06	0.10	N/A	0.08	0.10	N/A
Polarizability	Scalar Property	0.41	0.44	N/A	0.60	0.50	N/A
KDEEP [a]	Grid-based CNN	0.72	-- [b]	2.09	0.65	-- [b]	1.92
RF-Score [a]	RF Docking Score	0.78	-- [b]	1.99	0.75	-- [b]	1.66

Table 7.1: **Performance evaluation of models trained on PDBbind refined version 2016 dataset on unique complexes in the CSAR NRC-HiQ test sets.** Results reported as Pearson correlation coefficient (R), Spearman rank correlation coefficient (ρ), and root mean square error (RMSE). Note that the Spearman rank correlation here is across all targets in the coreset, while the "ranking power" metric is based on within-target ranking of molecule affinities.

Next, in the interest of obtaining a more complete benchmark and facilitating future comparisons, we extended our evaluation of the CSAR NRC HiQ test sets to the full molecule lists, which included 176 and 167 molecules in sets 1 and 2, respectively. Again, we re-trained our models on the PDBbind set, excluding now either the 176 or 167 compounds in test set 1 or 2 in addition to the remaining molecules in the 285 compounds from the coreset. Performance of BCL-AffinityNet on set 1 (R=0.75, ρ=0.75, RMSE=1.32) is very similar to performance on set 2 (R=0.74, ρ=0.73, RMSE=1.36) (Table 7.2).

		Benchmark Test Set					
		CSAR NRC-HiQ set 1 (n=176)			CSAR NRC-HiQ set 2 (n=167)		
Descriptor Set	**Model Type**	Pearson R	Spearman ρ	RMSE	Pearson R	Spearman ρ	RMSE
BCL-AffinityNet	DNN 2x512-32	0.75	0.75	1.32	0.74	0.73	1.36
Molecular Weight	Scalar Property	0.50	0.51	N/A	0.66	0.67	N/A
TPSA	Scalar Property	-0.03	0.08	N/A	0.28	0.27	N/A
LogP	Scalar Property	0.27	0.38	N/A	0.21	0.36	N/A
Polarizability	Scalar Property	0.56	0.59	N/A	0.68	0.69	N/A

Table 7.2: **Performance evaluation of models trained on PDBbind refined version 2016 dataset sans CSAR NRC-HiQ complexes on all complexes in the CSAR NRC-HiQ test sets.** Results reported as Pearson correlation coefficient (R), Spearman rank correlation coefficient (ρ), and root mean square error (RMSE). Note that the Spearman rank correlation here is across all targets in the coreset, while the "ranking power" metric is based on within-target ranking of molecule affinities.

7.2.8 Ranking power performance evaluation

The CASF2016 ranking power evaluation analyzes the ability of score functions to rank ligands targeting the same receptor. Among the methods originally compared in Su et al. 2019, BCL-AffinityNet (ρ=0.69) places just after ΔVinaRF¬20 (ρ=0.75) (Figure 7.4). Again taking into consideration Lu et al. 2019 re-training ΔVinaRF¬20 to exclude the 140 overlapped test set compounds, ΔVinaRF¬20 achieves a ranking power ρ=0.63 compared to ΔVinaXGB which achieves a ranking power of ρ=0.65 45.

Altogether results on the scoring power and ranking power tests suggest that BCL-AffinityNet is competitive with state-of-the-art SB virtual screening methods for binding affinity prediction and affinity ranking.

7.2.9 Docking power performance evaluation

Despite its success in the scoring and ranking power evaluations, BCL-AffinityNet is not ideally suited for decoy discrimination. This is because the training set for BCL-AffinityNet is composed entirely of native protein-ligand complexes. Thus, while BCL-AffinityNet could likely be used with an AD model generated in the same feature space to exclude clashed structures (by virtue of the lack of occupancy in the shortest distance bins, Figure 7.1A), it is unlikely to be able to discriminate plausible docking poses.

To address this limitation, we built a shallow multitasking ANN trained with the same PLC descriptors (eq. 2) as BCL-AffinityNet, with the addition of the hydrogen bond angle descriptor described above (see Supporting Information for a sample code object file). We reasoned that in differentiating properly docked poses it would be insufficient to consider only hydrogen bond donor/acceptor distances. In our experience,

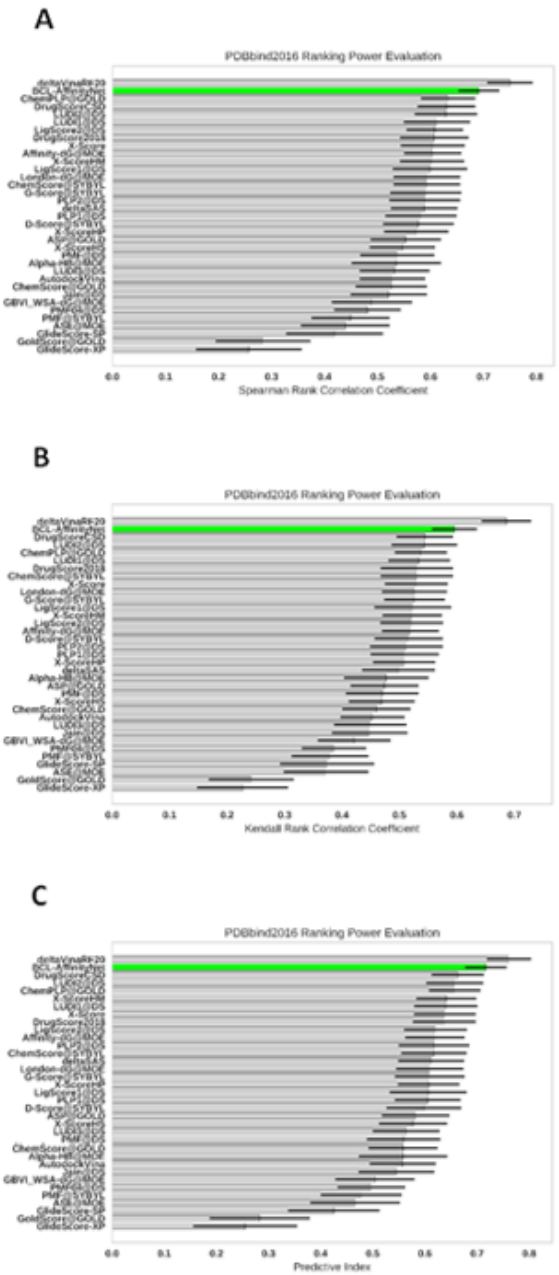


Figure 7.4: Ranking power evaluation of BCL-AffinityNet. Comparison of BCL-AffinityNet ranking power to other methods from the CASF2016 benchmark by Su et al.1 with (A) Spearman rank correlation coefficient, (B) Kendall rank correlation coefficient, and (C) predictive index. Error bars indicate the 90% confidence interval. Green bars indicates BCL-AffinityNet.

we have also found that separating a categorical prediction task (i.e. is this pose most likely to be 1.0 $\AA$ from the native pose, 2.0 $\AA$, or 5.0 $\AA$) into separate classification tasks for each category generally does not worsen model performance but may improve it. Thus, we nominally organized the output layer as five correlated classification tasks: determining whether a pose was less than 1.0 $\AA$, 2.0 $\AA$, 3.0 $\AA$, 5.0 $\AA$, and 8.0 $\AA$ from the native pose.

We trained this ANN on the PDBbind v.2016 refined set excluding all coreset protein-ligand complexes. For each complex in the training set, 250 additional decoys were generated with RosettaLigand (see Methods for details). The final model score, which we refer to as BCL-DockANNScore, is the product of the classification probability of a pose being less than 2.0 $\AA$ from the native pose and the BCL-AffinityNet affinity prediction score for that pose.

BCL-DockANNScore performs reasonably well on the docking power benchmark with success rates of 0.81, 0.91, and 0.95 for native pose recovery at a 2.0 $\AA$ threshold for poses within the best scoring 1, 2, and 3 poses, respectively (Figure 7.5). When native poses are excluded, BCL-DockANNScore success rates reduce by $\tilde{5}$%, consistent with performance reductions in multiple other methods (Figure S15). Binding funnel analysis of BCL-DockANNScore demonstrates good Spearman rank correlation coefficients at wide RMSD ranges, but performs less well in the 0 – 2.0 $\AA$ range (Figure S16). This suggests that one possible route to improve BCL-DockANNScore further is to provide additional training decoys within the 0 – 2.0 $\AA$ range or additional high-resolution descriptors.

7.2.10 Screening power performance evaluation

We evaluated BCL-DockANNScore on the forward and reverse screening tests. The forward screening power task evaluates the ability of a score function to identify small molecule ligands that bind to a target protein. The reverse screening power task evaluates the ability of a score function to identify the protein that most effectively binds a small molecule ligand (i.e. cross-docking)1.

Similar to the docking power evaluation, we find that BCL-DockANNScore performs reasonably well, but not always among the very best docking scores. On the forward screening task, BCL-DockANNScore has a success rate of 0.18, 0.33, and 0.58 when identifying the ligand amongst the top 1%, 5%, and 10% of candidates, respectfully (Figure 7.6A). This is competitive with the best score functions at the 10% level; however, performance at the 1% level is more mid-tier (ranking alongside several of the MOE score functions, while the top performers are from GOLD, Glide, and the AutoDock Vina and derived methods). The overall enhancement factor at the 1% level is 8.5 (Figure 7.6C). In contrast, we find that the performance on the reverse screening task is competitive even with the top-performers when identifying the top 1%, 5%, and 10% of candidates, with success rates of 0.15, 0.24, and 0.39, respectively (Figure 7.6B).

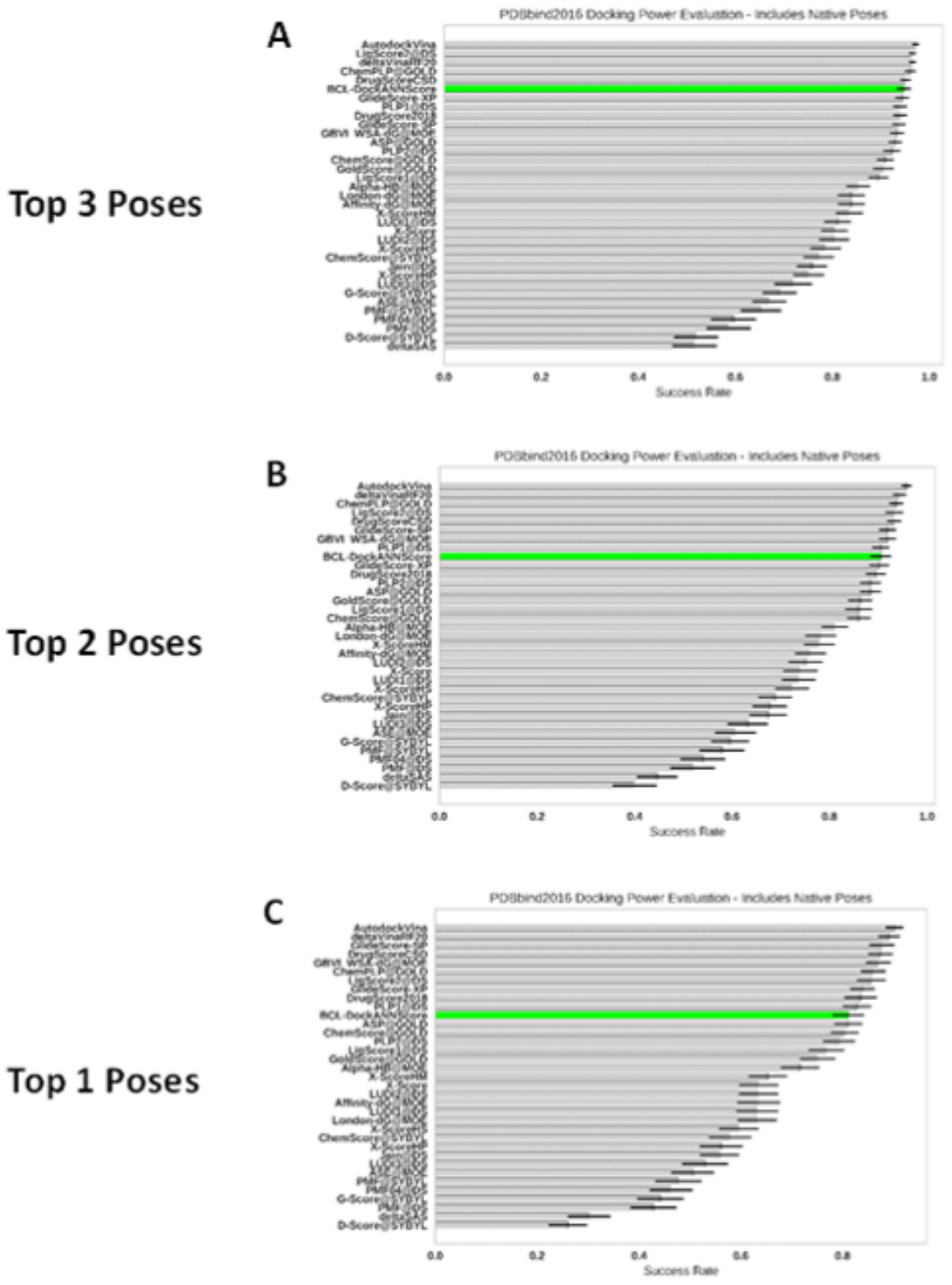


Figure 7.5: Docking power evaluation of BCL-DockANNScore. Comparison of BCL-DockANNScore docking power to other methods from the CASF2016 benchmark by Su et al.1 when recovering the native pose under 2.0 $\AA$ RMSD (A) within the top 3 poses, (B) within the top 2 poses, and (C) within the top 1 poses. Error bars indicate the 90% confidence interval. Green indicates BCL-DockANNScore.

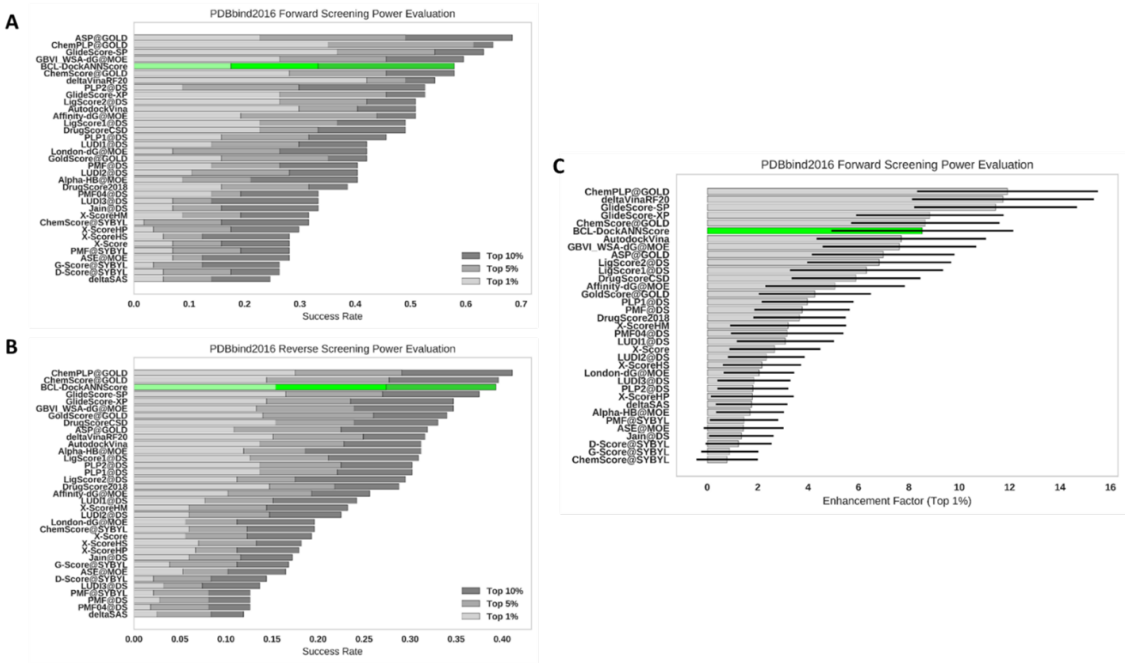


Figure 7.6: Screening power evaluation of BCL-DockANNScore. Comparison of BCL-DockANNScore screening power to other methods from the CASF2016 benchmark by Su et al.1. (A) Forward screening power evaluation success rates, (B) Reverse screening power evaluation success rates, (C) Forward screening power evaluation enhancement factor (top 1%). Error bars indicate the 90% confidence interval. Green indicates BCL-DockANNScore.

7.2.11 Generating absolute pharmacophore maps

Finally, one important consideration in the development of a SB score function for the BCL was model interpretability. One of the strengths of SB CADD is that predicted changes in activity can be attributed to specific interactions with the target. Neural networks are, however, often negatively characterized as "black boxes" because usually the function learned in the model cannot be decomposed into human interpretable parts. Traditional docking scoring functions, such as RosettaLigand, have the advantage that they can be decomposed into target per-residue contributions to the overall predicted affinity. This is important in drug discovery, where predictions need to be actionable. Here, we demonstrate that BCL-AffinityNet predictions can be decomposed into a map of atom contributions to the predicted bioactivity.

We take two general approaches for constructing a pharmacophore map: (1) Absolute feature contributions (Figure 7.7) and (2) relative feature contributions (Figure 7.8). The first case generates a map on any individual molecule by evaluating the contributions of specific atoms to the overall predicted activity. This can be likened to evaluating model input sensitivity, except in this case the molecule of interest is being perturbed instead of the weights connecting individual neurons in the model.

To generate an absolute pharmacophore map of a given molecule, we perturb the chemical structure by

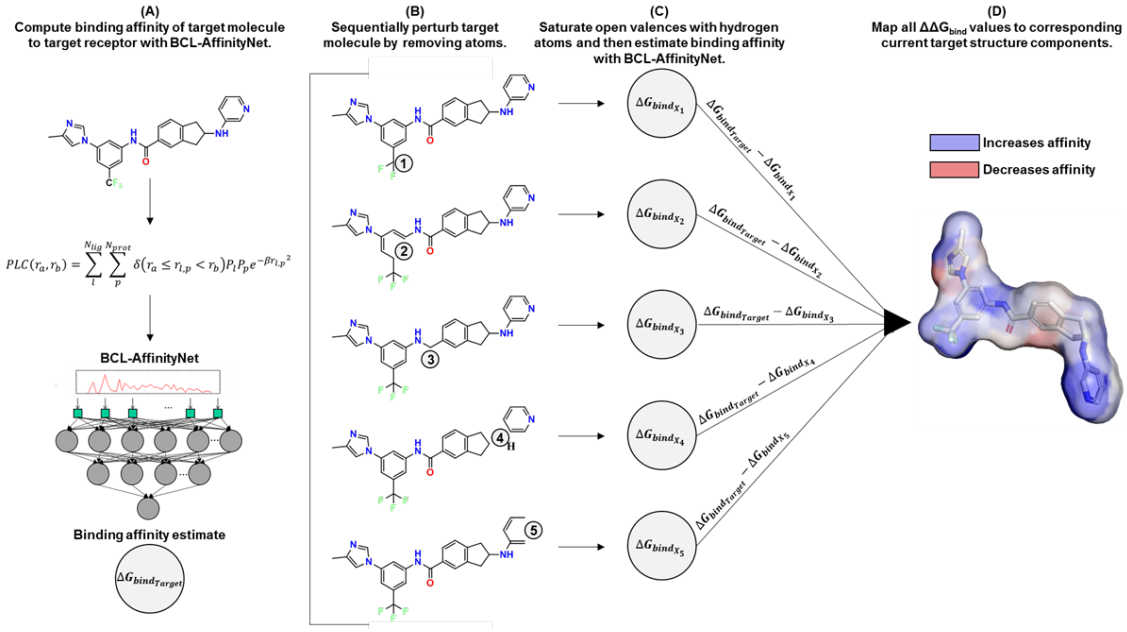


Figure 7.7: Construction of absolute pharmacophore maps. (A) The target molecule, in this case compound 7c from Zhu et al.54, is first modeled in complex with its target receptor using PLC descriptors and scored with BCL-AffinityNet. (B) Then we iterate over each atom in the target molecule and sequentially remove it from the molecule to create a perturbed molecule, X. (C) Perturbed molecules are saturated with hydrogen atoms to close any open valences resulting from the perturbation, and then they are scored with BCL-AffinityNet. (D) The differences in predicted binding affinity between the starting molecule and each perturbed molecule are mapped to the corresponding atoms of the starting structure. Here, predictions are in units of kcal/mol at 300K. The surface representation of atoms that contribute beneficially to BCL-AffinityNet's binding affinity prediction are blue, while atoms that worsen the prediction are in red. Atoms that contribute neutrally/negligibly are white.

sequentially removing individual atoms and closing the newly opened valence(s) with hydrogen atoms. Afterward, we compute the predicted affinity for each perturbed molecule with BCL-AffinityNet. The predicted binding affinity of the perturbed molecules are compared to that of the original molecule. The differences in predicted activity between the perturbed and original molecules are assigned to the corresponding atoms (Figure 7.7).

7.2.12 Generating relative pharmacophore maps

Relative pharmacophore maps leverage structural similarity within a congeneric ligand series to attribute predicted affinity differences to specific substructures. It has been shown that highly accurate binding affinity estimates can be obtained with alchemical free energy methods when reference structures with experimentally determined binding affinities within a congeneric ligand series are available (Wang et al., 2019a, 2015; Zou et al., 2019).

To generate a relative pharmacophore map between two molecules, we first identify a common substructure (MCS) via one of two methods: (1) identify the largest subgraph isomorphism between the two molecules, or (2) assign spatially mutually matched atoms to be common to one another (the first approach is more accurate and is the default approach). Component substructures that graphically correspond to the same common atoms are then iteratively removed, newly opened valences are closed with hydrogen atoms, and the perturbed molecules are scored with BCL-AffinityNet (Figure 7.8).

Thus, for each non-MCS substructure in the reference and target molecules there is a ΔΔG_bind between the non-perturbed and perturbed molecules. A final ΔΔΔG_bind is computed for each non-MCS substructure as the difference between the reference and target perturbation ΔΔG_bind values (Figure 7.8). The ΔΔΔG_bindvalues are mapped to the target molecule for visualization.

Consider a series of type II tyrosine kinase inhibitors (TKIs) of DDR1 kinase developed recently by Zhu et al (Zhu et al., 2019). We generated relative pharmacophore maps of compounds 7c, 7f, and 7j to compound 7i (Zhu et al., 2019) (Figure 7.9 A – D). We also modified the compound 7 scaffold to include NC mutations in the hinge-binding region analogous to prior substitutions done by Wang et al. (Wang et al., 2016) in a previous DDR1 TKI series (Figure 7.9 A, E – G).

From the pharmacophore maps, we also compute relative binding affinities of each molecule to compound 7i by summing the ΔΔΔG_bind values for each non-MCS component in the target molecule: ΔΔG_bind=Sum(ΔΔΔG_bind). In all comparisons, the trifluoromethyl group is preferable to the methyl. Relative binding affinity estimates of compounds 7c and 7f from 7i are within 0.50 kcal/mol of experimental values (-2.62 vs. -2.82 kcal/mol and -2.32 vs. -2.25 kcal/mol, respectively) (Figure 7.9 A, C – D). The ethyl in 7j is also correctly estimated to improve binding affinity relative to methyl in 7i; however, BCL-AffinityNet underestimates the extent of

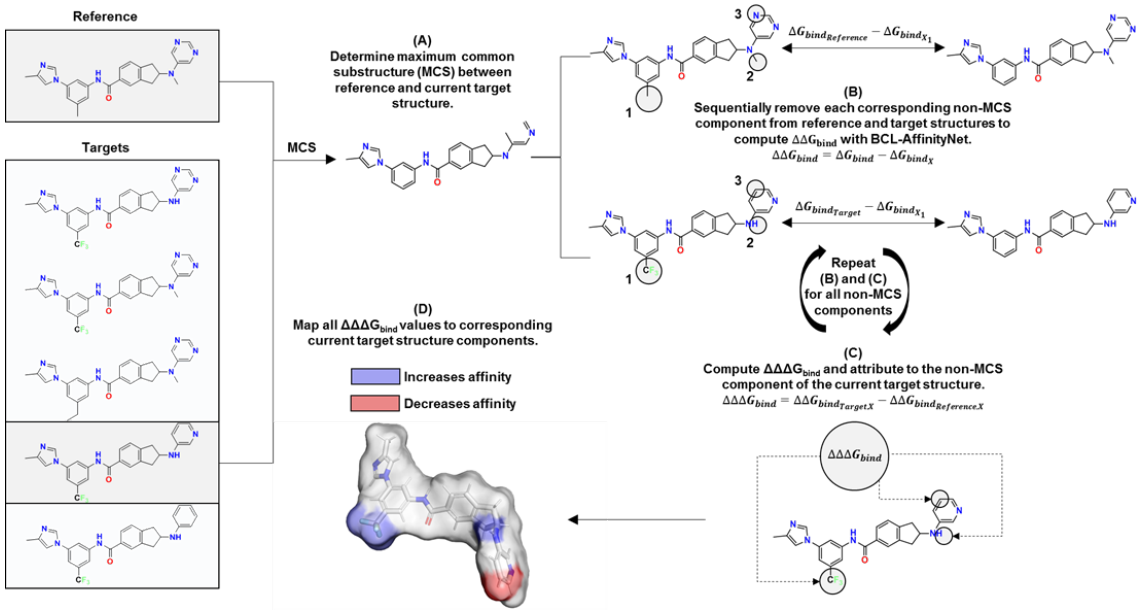


Figure 7.8: Construction of relative pharmacophore maps. Relative pharmacophore maps are generated from a target molecule and a reference molecule. (A) Determine the MCS between the reference and target structure. (B) Identify the MCS atoms that connect to corresponding non-MCS substructures in both the reference and target molecule. Non-MCS atoms are circled in grey and corresponding substructures between the reference and target share numerical labels (e.g. the reference molecule methyl circled in grey and the target molecule trifluoromethyl circled in grey are correspond structurally and are labeled "1"). For both the reference and target molecule, non-MCS substructures are independently removed. The binding affinities of the reference, target, and perturbed molecules are estimated with BCL-AffinityNet. The ddG_{bind} between starting and perturbed molecules is determined for both the reference and target. (C) For each corresponding non-MCS substructure, compute $dddG_{bind}$ as $ddG_{bind_{(Target,X)}} - ddG_{bind_{(Reference,X)}}$, where X indicates the perturbed target or reference molecule. (D) Map the $dddG_{bind}$ values back to the target molecule non-MCS substructures. The surface representation of atoms that contribute beneficially to BCL-AffinityNet's binding affinity prediction are blue, while atoms that worsen the prediction are in red. Atoms that contribute neutrally/negligibly are white.

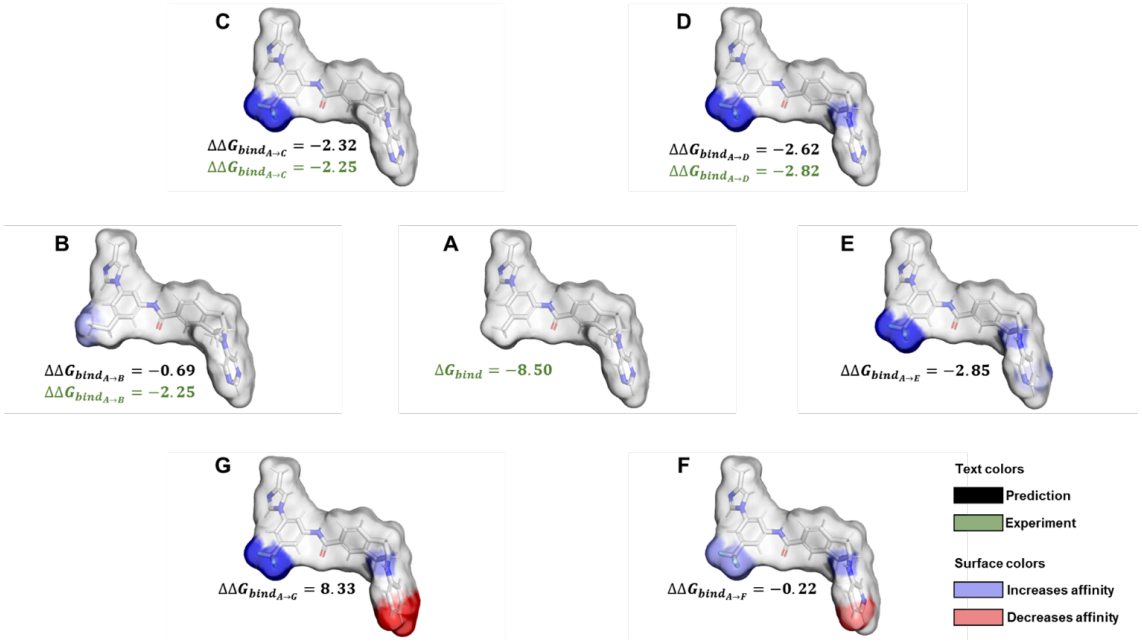


Figure 7.9: Figure 7.9. Relative pharmacophore maps of a congeneric DDR1 inhibitor series. (A) Compound 7i is the reference molecule for creation of the pharmacophore maps. Compounds (B) 7j, (C) 7f, and (D) 7c from Zhu et al.54. Compounds with the NC alteration at (F) the hinge-binding nitrogen atom, (E) the symmetrically placed hinge-binding nitrogen rotated away from the from the hydrogen bond donor partner, and (G) both nitrogen atom positions at the hinge-binding ring. Binding affinities in black text are predicted by BCL-AffinityNet, while green values are from Zhu et al. (Zhu et al., 2019). The surface representation of atoms that contribute beneficially to BCL-AffinityNet's binding affinity prediction are blue, while atoms that worsen the prediction are in red. Atoms that contribute neutrally/negligibly are white.

the affinity improvement (-0.69 vs. -2.25 kcal/mol) (Figure 7.9 A – B). Conversion of both hinge binding nitrogen atoms to carbon atoms is strongly unfavorable even in the presence of the trifluoromethyl group, consistent with prior SAR (Wang et al., 2016) (Figure 7.9 A, G). Thus, the relative pharmacophore maps provide meaningful QSAR insights that can be readily visualized.

Relative pharmacophore maps can be generated with respect to one or more reference input molecules (e.g. hit compounds or scaffolds), or in a pairwise fashion across a series of input molecules. If more than one molecule is used as a reference, the final map for each target molecule indicates the favorability of each molecule's substitutions in comparison to the whole ensemble. For an example command-line to generate a relative pharmacophore map, see the Supporting Information.

7.2.13 A case study on guiding chemical modifications with pharmacophore maps

To illustrate further this approach, consider three congeneric dysiherbaine analogs in complex with ionotropic glutamate receptor 5 (iGluR5). These molecules differ from one another by small substitutions at carbon atoms (1) and (2) (Figure 7.10A – C, first row). Each of the analogs was scored with BCL-AffinityNet and ranked correctly. For each of these three compounds, we generated absolute and relative pharmacophore maps (see Methods for command-line details).

First, we generated relative pharmacophore maps of the dysiherbaine analogs in the pairwise manner described above (Figure 7.8).The pharmacophore maps of dysiherbaine and neodysiherbaine suggest that the methylamine and hydroxyl substitutions, respectively, at position (2) provide a net increase in affinity relative to the proton in 8, 9-dideoxyneodysiherbaine (Figure 7.10A – C, third row). Furthermore, the pharmacophore maps predict that the methylamine modification increases binding affinity more than the hydroxyl substitution, in agreement with experimental observation (Figure 7.10B, C, third row).

Interestingly, the relative pharmacophore map of neodysiherbaine also predicts that the hydroxyl substitution at position (2) is more important for binding affinity than the hydroxyl substitution at position (1) (Figure 7.10B, third row). Similarly, the methylamine at position (2) of dysiherbaine is predicted to contribute more to the binding affinity than the hydroxyl at position (1) (Figure 7.10C, third row). Finally, we see from the absolute pharmacophore maps of all three analogs that the two carboxylic acid groups contribute favorably to binding. Indeed, we see that their contributions are predicted to be more important than the substitutions at (1) and (2), supporting the notion that these substituents are an important component of the conserved scaffold (Figure 7.10A – C, fourth row).

Together with the DDR1 TKI congeneric series, these comparisons illustrate how BCL-AffinityNet can yield structure-activity insight. To our knowledge, this is the first modern machine learning-based SB score function that is readily accompanied by an interpretable decomposition scheme. In principle, our pharma-

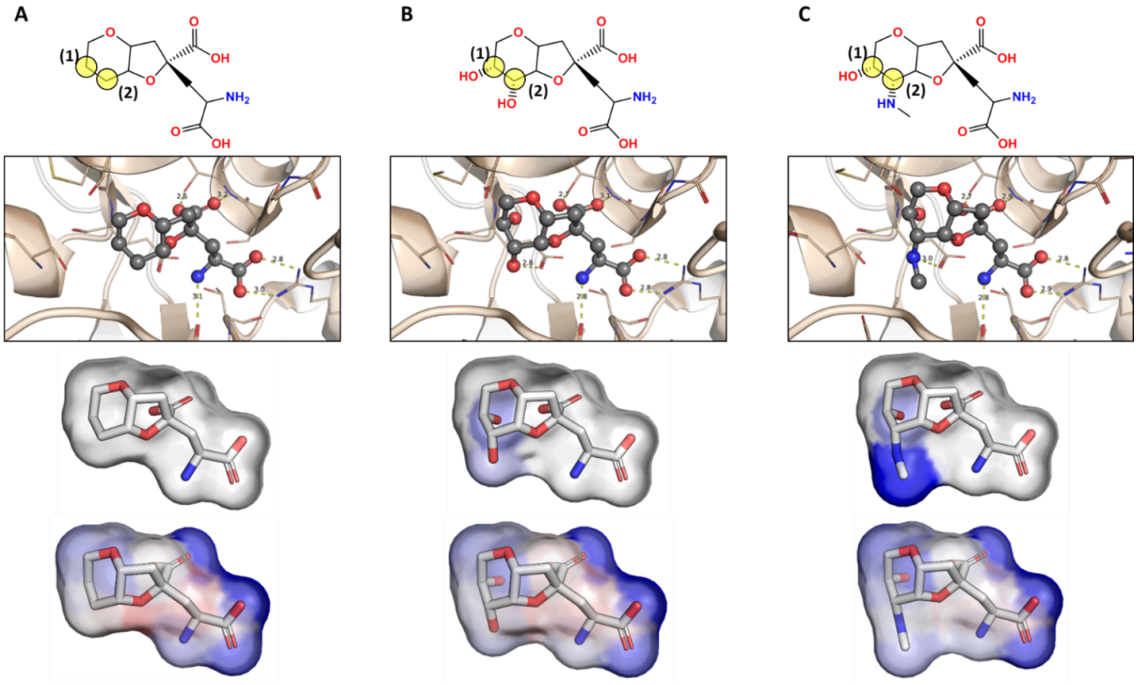


Figure 7.10: Pharmacophore maps of dysiherbaine analogs in complex with iGluR5 generated from BCL-AffinityNet. Pharmacophore maps were generated for iGluR5 complexed with (A) 8, 9-dideoxyneodysiherbaine (PDB ID 3GBB; pKd = 6.9, ΔG = -9.79 kcal/mol at 310K), (B) neodysiherbaine (PDB ID 3FV2; pKd = 8.1, ΔG = -11.49 kcal/mol at 310K), and (C) dysiherbaine (PDB ID 3FV1; pKd = 9.3, ΔG = -13.19 kcal/mol at 310K) and mapped onto the native bound pose. Labeled yellow transparent circles in top panel are used to reference the substituted carbon atoms of interest. Per atom pharmacophore map scores are output to a PyMol script for visualization as a molecular surface colored on a per atom basis by spectrum from blue (negative) to white (zero) to red (positive). In this example, negative values indicate atoms whose removal results in a loss in predicted binding affinity. The second row illustrates each ligand in complex with iGluR5. The third row illustrates the common substructure pharmacophore map (i.e. pairwise per-substructure relative binding free energy changes). The fourth row illustrates the raw pharmacophore map for each ligand upon sequentially removing individual atoms and saturating open valences.

cophore mapping procedure is compatible with any LB or SB machine learning score function in the BCL. Thus, these results demonstrate a fast and simple approach to generate interpretable pharmacophore maps from BCL machine learning model predictions.

7.3 Discussion

Here, we develop a novel machine learning-based score function for vHTS SB scoring. Our approach centers around the development of novel protein-ligand signed property correlation descriptors. In addition to the new descriptors, our models avoid the use of ligand-specific features to reduce the risk of training dataset bias. The new models, BCL-AffinityNet and BCL-DockANNScore, have been evaluated on current best practices benchmarks and compared to other standard and leading methods.

BCL-AffinityNet generally performs on par with or better than currently available SB virtual screening scores in affinity prediction and affinity ranking. BCL-DockANNScore, while generally not as good as GOLD, Glide, or the AutoDock Vina and derived methods at pose recovery or screening, performs competitively with respect to all of the evaluated methods. We therefore suggest that it may be a generally useful SB scoring algorithm with especially strong affinity prediction. Indeed, some of the best methods for docking and screening failed to provide estimates for power scoring (e.g. statistics for GlideScore-XP are based on 258/285 protein-ligand pairs, GlideScore-SP 252/285, GoldScore@GOLD 244/285)1. Thus, when considering all of the tasks together (scoring power, ranking power, docking power, and screening power), the new SB scoring models in the BCL demonstrate the utility of our novel signed property protein-ligand correlation descriptors for SB CADD. Moreover, BCL-AffinityNet and BCL-DockANNScore represent the first instantiation of SB scoring in the BCL.

While a number of algorithms consider multiple ligand-specific descriptors in their feature space alongside the protein-ligand interaction features (e.g. AutoDock Vina incorporates e.g. the ligand length, number of hydrophobic atoms, etc.; ΔVinaRF, and ΔVinaXGB both include ligand-specific pharmacophore features; ΔVinaXGB includes an estimate of ligand conformational stability; KDEEP contains ligand-specific voxels colored by pharmacophore features) (Su et al., 2019; Wang and Zhang, 2017; Jiménez et al., 2018; Lu et al., 2019), we made a conscious decision to avoid inclusion of such features in BCL-AffinityNet and BCL-DockANNScore. This was done to reduce the ligand bias of the models and hopefully yield a more generalizable score function. Nevertheless, efforts are underway to incorporate other aspects of protein-ligand binding affinity other than just interaction score terms into the BCL-AffinityNet and BCL-DockANNScore in an unbiased manner. These include improvements to both the neural network architectures employed here as well as incorporation of efficient metrics for solvation energy, ligand conformational preference, and entropy changes.

An important limitation of our work is that all models were trained in the absence of explicit water molecules, metal ions, and/or other cofactors. Others have recently demonstrated that incorporation of explicit water molecules can improve model performance (Lu et al., 2019), and future improvements to our model will incorporate these elements. As these updates are introduced, we will also continue to retrain the models leveraging the increasing availability of high quality protein-ligand co-crystal structures with Ki/Kd data.

Another limitation is the under-optimized protein-ligand interaction feature space of the current models. The generalizability of the PLC descriptors used to build BCL-AffinityNet and BCL-DockANNScore should not be conflated with completeness of the score function. By analogy, RosettaLigand with the Rosetta Talaris2014 score function (O'Meara et al., 2015) does not model halogen σ-hole interactions with aromatic ring systems and is thus unlikely to accurately determine the protein-ligand binding affinities of systems with these interactions. In the same way, BCL-AffinityNet and BCL-DockANNScore are incomplete representations of protein-ligand interactions. Further score function development will focus on expanding the availability of training data as well as describing additional salient chemical features.

Ongoing work in the Meiler Lab is focused on the development of both LB and SB small molecule de novo design and focused library design algorithms. A critical motivator for the present work was the need for the BCL to have a rapid and flexible SB score function that can be deployed for design tasks where there is insufficient data to build a reliable LB QSAR model. BCL-AffinityNet and BCL-DockANNScore are fully integrated into the BCL descriptor framework, allowing them to be called and mathematically combined with a multitude of other features, including AD scores, ligand descriptors, and more.

Another fundamental hurdle that we wanted to overcome was the so-called "black box" problem. This problem arises whenever the underlying feature space of the score function cannot be decomposed into human-interpretable parts, and it presents a major challenge when relying on complex score functions for rational drug design. In this manuscript, we have demonstrated a simple approach that can be employed with any score function in the BCL (machine learning or not) to convert predictions into all-atom pharmacophore maps. These pharmacophore maps can be generated with respect to underlying substructures or spatially matched atoms between different molecules, or they can be generated for individual molecules without a reference structure. We demonstrate how this can be accomplished with the BCL-AffinityNet score function for a series of congeneric DDR1 TKIs and dysiherbaine analogs. The relative pharmacophore maps provide an interpretable decomposition of affinity with respect to scaffold modifications that can be used to guide further molecule optimization. The absolute pharmacophore map procedure can tell the user which atoms are most salient to BCL-AffinityNet's predictions. In addition to being a useful tool for interpreting machine learning score functions in the BCL, we anticipate that such pharmacophore maps will be valuable in automated drug

design tasks.

All of our models and applications for generating new models are freely available with an academic license for the BCL at http://meilerlab.org/. We hope that our descriptors and models may be integrated with future machine learning architecture development and descriptor optimization for the continued advancement of drug discovery.

7.4 Methods

7.4.1 Training dataset preparation

BCL-AffinityNet was trained using the refined set plus protein-ligand complexes from the general set of the PDBbind v.2016 dataset that satisfied the following criteria: (1) the ligand was a small molecule; (2) the binding affinity was measured as either Ki or Kd; (3) all atom types had defined Gasteiger atom types. The PDBbind v.2016 coreset was not included in the training set for any of the models for any of the performance evaluations. BCL-DockANNScore was trained using the refined set protein-ligand complexes from the PDBbind v.2016 dataset excluding the 285 coreset compounds. For each protein-ligand complex in the PDBbind v.2016 refined set, 250 additional pose decoys were generated with RosettaLigand flexible docking with the Talaris2014 score function (Fu and Meiler, 2018; Smith and Meiler, 2020; Meiler and Baker, 2006).

7.4.2 Model validation

Metrics for scoring power, ranking power, docking power, screening power, and confidence interval bootstrapping were performed with the scripts made available with download of PDBbind v.2016.1 All models were trained with five-fold random-split cross-validation. The final model prediction value is the average prediction value obtained across all five splits (i.e. as opposed to selecting a single best model from the five splits). PDBbind v.2016 coreset complexes were always excluded from training. For other external test-set evaluations, the models were always re-trained excluding all test-set complexes explicitly. Thus, the final training set sizes for testing on the PDBbind 2016 coreset (n=285), CSAR NRC-HiQ 1 Jimenez et al. subset (n=55), CSAR NRC-HiQ 2 Jimenez et al. subset (n=49), CSAR NRC-HiQ 1 full set (n=176), and CSAR NRC-HiQ 2 full set (n=167) were 7568, 7551, 7537, 7442, and 7440 (not every complex in the CSAR sets is in the PDBbind v.2016 set, hence the differences are not equivalent to 7568 - n). For comparisons to the CSAR NRC-HiQ benchmarks in Jimenez et al., 201832, complexes present in both the CSAR test sets and the PDBbind v.2016 refined subset were removed from the CSAR test sets. This resulted in two CSAR test sets of sizes 55 and 49, respectively, with the exact same PDB IDs as reported in the supplemental material of Jimenez et al., 2018.

For our baseline assessment of ligand and receptor pocket bias on the PDBbind v.2016 coreset, we trained

two DNNs identical in architecture to BCL-AffinityNet. For descriptors, we utilized the same chemical features, distance bins, and sign pairings as in the PLC descriptors, except we instead generated signed 3D autocorrelations of the ligand and/or receptor itself15. As inputs, we used the structures provided in the PDBbind v.2016 dataset such that the ligand-based DNNs were trained on the native poses of the ligands and the pocket-based DNNs were trained on the receptor binding pockets as extracted for inclusion in PDBbind v.20161, (Liu et al., 2015).

For validation splits that explicitly address ligand and pocket bias of the training datasets, we generated k-means (k=75) AD models of the PDBbind v.2016 coreset (n=285) based on ligand 3DAs, pocket 3DAs, or column-combined ligand and pocket 3DAs (using the same descriptors that were used to create ligand- and pocket-based QSAR models; see Supporting Information). We then scored all 7568 training set samples with each of these AD models. Previous studies on appropriate cutoffs for distance-based AD models have suggested that test set samples further away from their closest node than 95 – 100% of the training samples can reliably be considered outside of the domain of applicability (Minovski et al., 2013; Sahigara et al., 2012). We therefore made three test-set splits (one for each AD model) containing all training samples that had AD scores greater than 1.0. The resulting test sets are those samples whose ligands, proteins, or ligands and proteins can be considered within the same AD as the PDBbind v.2016 coreset. Put another way, this creates larger PDBbind v.2016 coreset-like leave-class-out test set splits based on the properties of the ligands, protein pockets, or combined ligands and protein pockets. We refer to these test sets respectively as LBAD test (n=995), pocket AD test (n=379), and combined AD test (n=1377). For these evaluations, total model training sample size is 7568 - n. For details on command-line syntax, see Supporting Information.

7.4.3 Training neural networks for affinity prediction and pose discrimination

All neural networks were trained with the BCL. Our binding affinity prediction model, which we call BCL-AffinityNet, is a single-task, feed-forward regression neural network trained to predict pKi/d. While technically a "deep" neural network in that we utilize two hidden layers (512 and 32 neurons, respectively) instead of just one, BCL-AffinityNet is quite small compared to neural networks recently published for similar tasks (Ragoza et al., 2017; Jiménez et al., 2018; Izhar Wallach, 2015). Our pose prediction model, which we call BCL-DockANNScore, is a shallow (single hidden layer, 32 neurons) multi-tasking feed-forward classification neural network that predicts whether a protein-ligand pose is less than 1.0, 2.0, 3.0, 5.0, and 8.0 $\AA$ from the correct pose. Both networks can thus be formalized as follows:

For a network with L hidden layers indexed lε(1..L), forward propagation for lε(0..L-1) can be described as

$$\boldsymbol{z}^{(l+1)} = \boldsymbol{w}^{(l+1)}\boldsymbol{y}^{l} + \boldsymbol{b}^{(l+1)}, \tag{7.3}$$

$$\boldsymbol{y}^{(l+1)} = f\left(\boldsymbol{z}^{(l+1)}\right), \tag{7.4}$$

where $\mathbf{y}^l$ is the output vector at layer l connected to the input vector $\mathbf{z}^{(l+1)}$ at layer $l+1$ by weights $\mathbf{w}$ and biases $\mathbf{b}$, and f is the transfer function applied to each set of inputs into the $l+1$ layer. Correspondingly, the activation of a single neuron i in hidden layer $l+1$ can be represented as

$$z_i^{(l+1)} = \boldsymbol{w_i}^{(l+1)}\boldsymbol{y}^{l} + b_i^{(l+1)}, \tag{7.5}$$

$$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right) \tag{7.6}$$

to yield the output $y_i^{(l+1)}$ from layer $l+1$. A mean-squared error (MSE) cost function was employed in all studies. Overtraining is prevented through the use of dropout in the input and hidden layers. During forward propagation each output value y_i^l of each i neuron in the layer l of the ANN is randomly multiplied either by a value of 0 (corresponding to a "dropped" neuron) or 1.

$$z_i^{(l+1)} = \boldsymbol{w}^{(l+1)}\left(\boldsymbol{r}^{l} * \boldsymbol{y}^{l}\right) + b_i^{(l+1)}, \tag{7.7}$$

Here, $\mathbf{r}^l$ is a vector with the same dimensions as $\mathbf{y}^l$ whose values are either 0 (at fraction p) or 1 (at fraction 1 – p). At the end of every training batch, $\mathbf{r}^l$ is shuffled. At test time the corresponding weights are scaled down by the factor 1–p.

The BCL-AffinityNet DNN contains two hidden layers with 512 and 32 neurons, respectively. It was trained with 5% dropout in the input layer, 25% dropout in the first hidden layer, and 5% dropout in the second hidden layer (Mendenhall and Meiler, 2016). All neurons utilized a leaky rectifier transfer function:

$$f(x) = \begin{cases} x & x > 0 \\ 0.05x & x \leq 0 \end{cases} \tag{7.8}$$

where x is the total input to a neuron. We utilized normalized mean absolute error (NMAE; defined as

the quotient of mean absolute error and mean absolute deviation) as our objective function during training.

The BCL-DockANNScore ANN contained a single hidden layer with 32 neurons. It was trained with 5% dropout in the input layer and 25% dropout in the hidden layer 6. All neurons utilized a sigmoid transfer function:

$$f(x) = \frac{e^x}{e^x + 1} \tag{7.9}$$

where x is the total input to a neuron. We utilized area under the curve (AUC) as our objective function during training.

The AffinityRF random forest model was trained with WEKA v.3.8.4 utilizing default settings.

7.4.4 Feature parameter and neural network hyperparameter tuning

Our adoption of 5% input layer dropout and 25% dropout in the first hidden layer (for both models) as well as the selection of a 32 neuron hidden layer prior to the output layer is based on extensive prior evaluation in Mendenhall et al. 2016. For classification models, it has been shown that shallow networks often perform equivalently and sometimes better than deep networks at a substantially reduced training cost5. This, coupled with our own experience with QSAR classification tasks, led us to use our previously utilized single hidden layer architecture for BCL-DockANNScore (Mendenhall and Meiler, 2016).

With respect to BCL-AffinityNet, we nominally selected the nearest power of 2 (29 = 512) to our input feature size as an upper limit for our first hidden layer size. We investigated two PLC descriptor feature parameters using five-fold random-split cross-validation with the DNN of this size: (1) the interaction bin distance, and (2) the smoothing parameter β (eq. 2). We selected an initial smoothing parameter value of 5.0 based on prior 3DA QSAR investigations in which values greater than one were effective (Mendenhall and Meiler, 2016).

Subsequently, we varied the interaction bin distances at 1.0 $\AA$ intervals between 4.0 and 9.0 $\AA$ and compared NMAE and Pearson correlation across the cross-validation splits. Our results suggested that distances greater than 5.0 $\AA$ were best (Figure S17). In the interest of keeping our feature set relatively small, we selected 7.0 $\AA$ for our final models. Similarly, we varied smoothing parameter between 0.1 and 10.0 at a fixed bin distance of 7.0 $\AA$. We found that β values between 3.0 and 10.0 produced similar results (Figure S18), so we retained a value of 5.0 for all additional studies.

With the PLC parameters selected, we then performed additional five-fold random-split cross-validation studies to determine an appropriate first hidden layer size. We decreased the number of neurons from 512 by powers of 2 down to the size of the second hidden layer (32 neurons). For completeness, we also evaluated a shallow ANN ranging in size from 32 – 128 neurons using either a leaky rectifier (eq. 8) or sigmoid (eq.

9) transfer function. Generally, we observed that shallow and deep networks with smaller (32 – 64 neurons) first hidden layers performed the worst independent of transfer function. We also noted that two hidden layers seemed better than one, with little improvement in cross-validation performance between 256 and 512 neurons (Figure S19).

We note that all cross-validation studies for PLC feature parameter and model hyperparameter tuning were done with the BCL-AffinityNet training set of size 7568 protein-ligand complexes (PDBbind v.2016 refined set excluding the coreset and including select general set complexes; see Methods subsection Model validation for details). Model performance on the external test sets was not evaluated during feature parameter or model hyperparameter tuning.

7.4.5 Resolving hydrogen bond angles in feature space

BCL-DockANNScore contains an additional feature type not present in BCL-AffinityNet. Specifically, we binned hydrogen bonding pairs by both distance and angle. We considered that the strength of hydrogen bonding interactions is often approximated not only with distances between donor and acceptor atoms but also with orientation angle. Therefore, we also developed a complementary feature to (eq. 2) to assist with the description of well-formed hydrogen bonds. While (eq. 2) is generalizable to any atom-based descriptor (or pair of descriptors if performing an asymmetric correlation) returning a scalar value, this descriptor is exclusively for hydrogen bond donor/acceptor pairs. Essentially, each distance interval specified by the boundaries r_a and r_b in (eq. 2) is equally partitioned into a user-specified number of bins (for this manuscript, nominally 45 bins of 8° each). Thus, for each distance bin there is also an angular component. See the Supporting Information for sample BCL code object files containing all properties employed in this study.

7.4.6 Input sensitivity analysis

The predictions for BCL-AffinityNet (and separately, BCL-DockANNScore) are the average predictions of the five cross-validated models. We can readily calculate feature importance for a single ANN by computing the magnitude of the input sensitivity across a dataset with respect to a given feature, after appropriate rescaling of the inputs. For model ensembles, the magnitude cannot be used or meaningfully averaged because feature input sensitivity may differ in sign for various feature-instance pairings. While we could look at the raw average of input sensitivity of models across a given instance-feature pairing, and then average the absolute value of that over the dataset, we suffer an issue with relative scaling of the input sensitivities, due to the non-linearity of the ANN's transfer function. Rather than deriving an optimal weighted feature importance metric for ANN ensembles by some criteria, we chose to simply evaluate how often the models in the ensemble agreed on the sign of the derivative for each feature, averaged across the dataset.

This is a form of input sensitivity analysis we refer to as "consistency". Here, we specifically evaluate the consistency of feature column perturbations on result labels across cross-validation models. Features for which models in the ensemble agree on the derivative sign most routinely are interpreted as those that are of most importance to the ensemble's performance. Consistency is thus insensitive to the magnitude of feature's influence.

To calculate consistency, we iterate across all input feature columns of a training sample, perturb the feature value by a small amount (e.g. 0.01), propagate the perturbed inputs, and measure the result. For efficiency, we perform a forward propagation pass, followed by a backpropagation pass with a slightly modified result, which is readily transformed into the forward input sensitivities. This is done for each cross-validation model (in this manuscript we performed five-fold cross-validation for all models). For each feature column, we count the number of models that predict that the perturbation will improve the score vs. the number of models that predict that the perturbation will worsen the score. This number is normalized such that when half of the models predict a negative change to the result and the other half predicts a positive change to the result the net consistency is zero. The consistency result is averaged across all examples in the training set for each individual feature.

CHAPTER 8

Simultaneous protein interface and small molecule design with BCL-Rosetta

This chapter is a collaborative work of Benjamin P. Brown, Jeffrey Mendenhall, Rocco Moretti, Sergey Lyskov, Alexander R. Geanes, Darwin Fu, Sandeep Kothiwale, Edward W. Lowe Jr., and Jens Meiler. This chapter is under review as a Brief Communication.

8.1 Introduction

Computer-aided drug design (CADD) has become a core component of modern drug discovery (Macalino et al., 2015). Both ligand-based quantitative structure-activity relationship (QSAR) modeling and structure-based docking virtual high-throughput screening (vHTS) have come-of-age as powerful tools for small molecule hit discovery (Geanes et al., 2016; Butkiewicz et al., 2013; Stein et al., 2020; DeLuca et al., 2015; Fu and Meiler, 2018; Meiler and Baker, 2006; Sadybekov et al., 2022). Advances in molecular mechanics methods such as free energy perturbation (FEP) and thermodynamic integration (TI) have led to unprecedented in silico rank-ordering of scaffold derivatives during hit-to-lead optimization (Wang et al., 2015; Zou et al., 2019). Ongoing investigations in machine learning (ML) and quantum chemistry are poised to increase the predictive power of our CADD score functions (Lu et al., 2019; Brown et al., 2021; Kirkpatrick et al., 2021; Gentile et al., 2020).

Despite these advances, there remains substantial attrition in the development of a compound from lead to FDA-approved therapy (Harrison, 2016; Waring et al., 2015). The time and cost required to develop a new drug remain substantial. Several algorithms have emerged that leverage the one-shot synthetic accessibility of made-on-demand libraries to propose efficient routes for molecular design (Sadybekov et al., 2022; Bellmann et al., 2022). Other algorithms leverage ML, combinatorial chemistry, and/or reaction-based design to generate small molecule libraries with favorable predicted properties and activities (Zhavoronkov et al., 2019; Brown et al., 2022).

All of these approaches represent important fundamental advances; however, they exist largely in isolation as highly specialized protocols. CADD requires adaptability. The nature and scope of a CADD challenge is heavily influenced by factors such as the availability of training data, knowledge of the target chemical space, the presence (or absence) of experimental characterization of the drug target and putative binding pocket(s), the flexibility (dynamics) of the target, the size of the system under investigation, the expected accuracy of the score function in the given system, and more. Thus, while specialized tools can be highly valuable in some circumstances, they may be of limited utility in others.

Here, we present a new modular, customizable framework for CADD that integrates the BioChemical Library (BCL) cheminformatics toolkit (Brown et al., 2022) with the Rosetta macromolecular modeling and design software suite (Leman et al., 2020). In addition to integrating the BCL into the broader Rosetta codebase, we have developed small molecule drug design "mutates", or chemical perturbations, with which to design new molecules. These mutates are capable of performing one-shot made-on-demand style reactions, single- and multi-component reactions, or reaction-free medicinal chemistry-inspired "alchemical" perturbations. The mutates are encoded as "Movers" in Rosetta, which means they can be recombined in a protocol-specific manner. Collectively, the mutates provide an avenue for highly customizable ligand- or structure-based drug design protocols using the RosettaScripts, PyRosetta, or standard BCL command-line API.

The new drug design tools can be seamlessly combined with the diverse repertoire of modeling tools in Rosetta. Rosetta has been developed to model and design atypical moieties such as non-canonical amino acids (Renfrew et al., 2012), post-translational modifications (Labonte et al., 2017), nucleic acids (Alford et al., 2017), membranes (Leman et al., 2015), and more. Rosetta also contains multiple tools for sampling small and large protein conformational changes (Leman et al., 2020). In this manuscript, we demonstrate how the BCL-Rosetta integration can be utilized to build custom protocols for tasks such as induced-fit drug design, chemogenetics drug design, and selectivity design.

8.2 Results

8.2.1 Customization of atom selections during drug design

The drug design "mutates", or chemical perturbations, are capable of performing one-shot made-on-demand style reactions (Figure 8.1A), single- and multi-component reactions (Figure 8.1B), or reaction-free medicinal chemistry-inspired "alchemical" perturbations (Figure 8.1C).

These mutates are encoded as "Movers" in Rosetta and can be modularly recombined to build protocols. Each Mover defines a perturbation type, a set of mutable atom specifications, a druglikeness filter from the BCL descriptor framework10, and potentially several perturbation-specific options. By default, all atoms are mutable, and an atom is randomly selected for perturbation by its mutate Mover. The subset of mutable atoms is refined through user-specification.

To illustrate the atom selection ability, we apply a series of three mutates to a 2-amino-8-methyl-6-phenylpyrido[2,3-d]pyrimidin-7-one scaffold tyrosine kinase inhibitor (TKI) scaffold (Okram et al., 2006) to create a type II TKI topology. First, we restrict the mutable atom selection to the hydrogen atom bonded to the carbon at index 13 and apply Alchemy to transform it into a carbon. Second, we restrict our mutable atoms to the complement of the common subgraph between our original molecule (0) and our new molecule (1) and allow the AddMedChem mutate to append an ethylamide. Third, we restrict our mutable atoms as

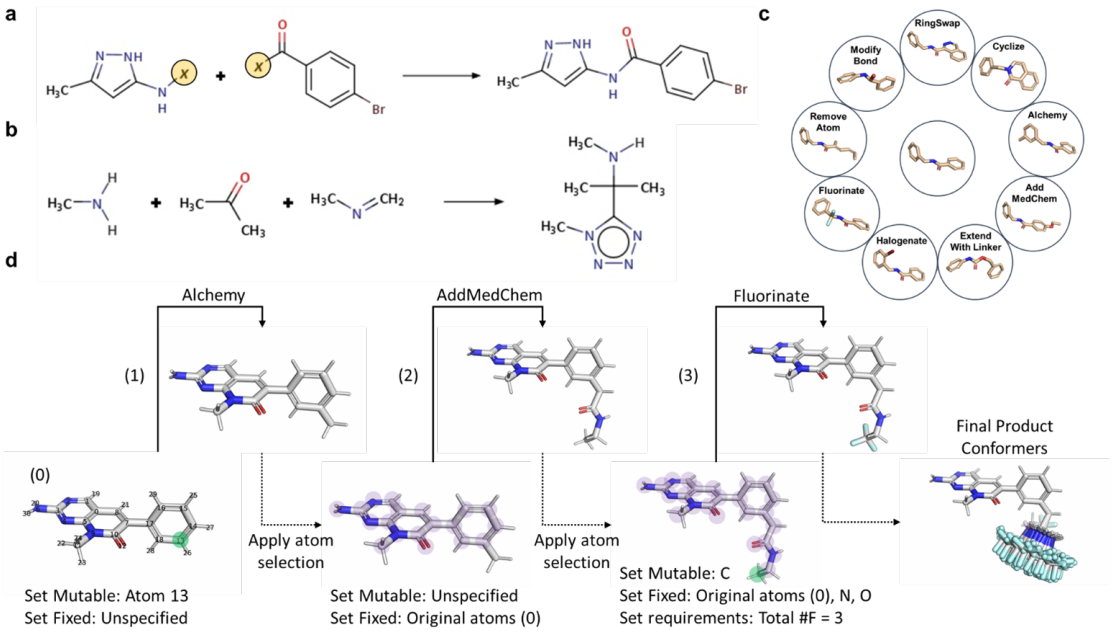


Figure 8.1: A modular framework for small molecule drug design chemical perturbations. a, One-shot chemical synthesis can be simulated by combining fragments with the AddMedChem mutate. The connection between the fragments is made through bonds at undefined atom types ("X"; yellow circles in reaction). b, Single- and multi-component reaction simulations can be performed with the React mutate. The reaction is read from an MDL RXN file where the product atoms are mapped to reagent atoms. c, Medicinal chemistry-inspired "alchemical" mutates can be performed without specifying chemical reaction pathways. d, Alchemical mutations allow user-specified restrictions on mutable (green circles) and fixed (purple circles) atoms.

previously, but also prevent design on heteroatoms and enable a Fluorinate-specific perturbation requirement that upon saturation the selected atom must have three attached fluorine atoms (see Online Methods). In Rosetta structure-based design protocols, by default conformer generation occurs only on the newly designed substructure. Notably, the mutates preserve coordinate information of the unperturbed substructure when possible (Figure 8.1D).

8.2.2 Illustration of induced-fit drug design in two receptors

We demonstrate a structure-based induced-fit drug design protocol to build TKIs that adhere to either a type I or II topology in the Abl kinase domain (KD) ATP binding pocket. Following guided stochastic design, TKIs are scored on low-resolution grids of 16 unique Abl kinase conformers (Meng et al., 2018). The best scoring complex is taken for high-resolution refinement and interaction energy evaluation. Using this approach, we observed an enrichment in activation loop (A-loop) inward (inactive) conformers with type II TKIs and A-loop outward (active) conformers with type I TKIs (Figure 8.2A).

We also demonstrate an induced-fit drug design simulation targeting a cryptic pocket in mAChR1 (Hollingsworth et al., 2019) that leverages Monte Carlo Metropolis (MCM) sampling of neighboring sidechains (Figure 8.2B). The open state of the cryptic pocket is enriched among designs that bury in the cryptic pocket, while the closed cryptic pocket state is enriched in ligands that are distant from the cryptic pocket (Figure 8.2C).

Finally, we show how the BCL-Rosetta integration can simulate "bump-and-hole" drug design for chemogenetics (Runcie et al., 2018). The strategy is to create a "hole" by mutating bulky receptor residues into smaller residues and then fill it by "bumping", or appending, atoms to a scaffold molecule (Figure 8.2D). We combine Rosetta protein sequence design with our drug design mutates to simulate experiments done by Runcie and colleagues to identify a mutant-ligand pair for BRD216. The correlation we obtain between computational and experimental (Runcie et al., 2018) relative binding affinities of BRD2 wild-type and L383V is consistent with recent RosettaLigand benchmarks (Smith and Meiler, 2020) (Figure 8.2E).

We expand on the conservative pocket redesign in Runcie et al. (Runcie et al., 2018) by enabling more sequence positions for design in our resfile (Methods). In a representative example of such a simulation, we obtain a "bump" modification that is a propylene, and a "hole" modification that is a combination of L41V and Y86V. The simulation also redesigned three additional residues (H91N, N87S, C83S) to increase hydrogen bonding of the ligand to the receptor (Figure 8.2F). This approach has clear applications for in silico design of DREADDs (designer receptors exclusively activated by designer drugs), which are a powerful means of noninvasively modulating cellular activity (Urban and Roth, 2015).

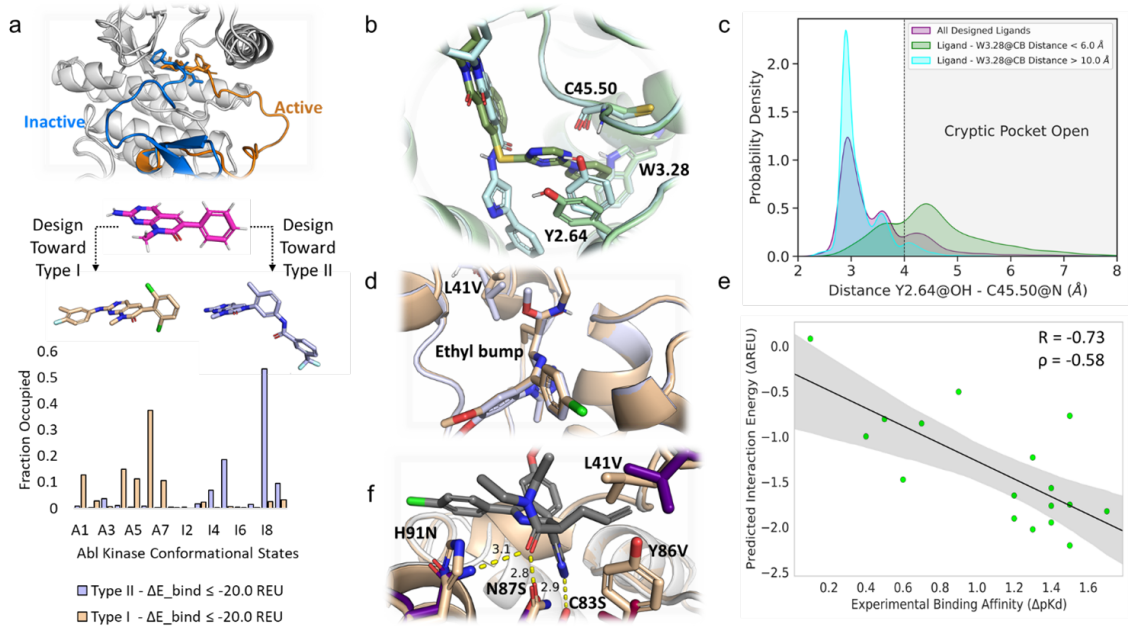


Figure 8.2: Small molecule design simulations can be performed in the presence of conformational changes and sequence design. a, Induced-fit design simulations of Type I or Type II tyrosine kinase inhibitors for Abl kinase captures activation loop conformational preferences. Design simulations were initiated with a common scaffold (magenta). Chemical perturbations of the scaffold were performed to generate either Type I (light brown) or Type II (light blue) inhibitors. b, Sample in silico designs that either do (light green) or do not (light blue) occupy the cryptic pocket of mAChR1. Protein colors match their corresponding ligands. c, Induced-fit design simulations of positive allosteric modulators (PAMs) targeting a cryptic pocket in muscarinic acetylcholine receptor 1 (mAChR1). The distance between the Y2.64 hydroxyl and C45.50 backbone nitrogen defines the accessibility of the cryptic pocket (Hollingsworth et al. 2019). d, Schematic representation of BRD2 bump-and-hole chemogenetic design simulation. BRD2 L41V mutation (light brown) is superimposed with wild-type (light blue-white). The "bump" corresponds to the ethyl (light brown) and the "hole" the reduction in size of L41 (light blue-white) to valine (light brown). e, Correlation between experimental (x-axis; Runcie et al. 2018) and computational (y-axis) relative binding affinity estimates between BRD2 and BRD2-L41V f, Simulating the simultaneous redesign of the BRD2 binding pocket and inhibitor scaffold.

8.3 Discussion

Rosetta has been developed to model macromolecule conformational changes and design with many polymeric species, such as non-canonical amino acids (Renfrew et al., 2012), post-translational modifications (Labonte et al., 2017),nucleic acids (Alford et al., 2017), membranes (Leman et al., 2015). Integration with the BCL extends this toolkit to enable versatile drug design simulations. We anticipate that this will be a valuable addition to the expanding array of tools available to computational chemists.

8.4 Methods

8.4.1 BCL command line syntax used to append an amide-linked trifluoroethyl group to a scaffold

The 2-amino-8-methyl-6-phenylpyrido[2,3-d]pyrimidin-7-one scaffold was obtained by manually removing the 3-(trifluoromethyl)benzamide and the core benzene ring methyl-substitution from the crystallized type II tyrosine kinase inhibitor (TKI) in PDB ID 2HIW12. The resulting scaffold molecule, saved as "LIG.sdf", was processed with the BCL molecule:Filter application to add hydrogen atoms, neutralize formal charges, and verify defined Gasteiger atom types.

```
bcl.exe molecule:Filter \
-input_filenames LIG.sdf -output_matched LIG.clean.sdf \
-add_h -neutralize -defined_atom_types
```

We visualized the molecule in PyMOL to identify the 0-indexed integer in "LIG.clean.sdf" corresponding to the carbon originally bonded to the 3-(trifluoromethyl)benzamide group. This atom (index 13) was selected as the atom from which to initialize our perturbation. The molecule:Mutate application was used to chemically perturb the scaffold "LIG.clean.sdf" into a type II TKI with the following command:

```
bcl.exe molecule:Mutate -input_filenames LIG.clean.sdf -output LIG.
    amide_cf3.sdf -implementation \
"Alchemy(mutable_atoms=13,allowed_elements=C,restrict_to_bonded_h=1)" "
    AddMedChem(mutable_fragments=LIG.clean.sdf,
    complement_mutable_fragments=1,medchem_library=/home/ben/workspace/
    bcl/rotamer_library/medchem_fragments/ethylamide.sdf.gz)" \
"Fluorinate(mutable_fragments=LIG.clean.sdf,complement_mutable_fragments
    =1,fixed_elements=N O, n_min_f=3,n_min_h_sub=3)" -random_seed
```

[breaklines]

If using the BCL mutates in Rosetta, the default conformational ensemble of "LIG.amide_cf3.sdf" will be generated by only sampling the new dihedrals added with the mutates. In the above BCL standalone

command, the final output is the scaffold with a single random conformer generated for the new atoms. To generate a conformational ensemble with BCL standalone with or without dihedral angle restrictions, see previously published protocols (Mendenhall et al., 2021; Brown et al., 2022).

8.4.2 Induced-fit drug design of type I and II TKIs in the Abl kinase domain

Each design was scored in 16 conformations of Abl kinase (taken from (Meng et al., 2018); 7 active and 9 inactive states annotated based on the activation loop being in the outward/active or inward/inactive conformation) using the Rosetta low resolution Transform mover. High resolution refinement was performed on the best scoring Abl kinase conformer with the designed ligand prior to computation of the interaction energy. The fraction of final protein-ligand complexes in each of the 16 conformational states was determined for either Type I (light brown) or Type II (light blue) design simulations. An equal number of simulations were initialized from each conformation for each inhibitor type.

All design simulations were run with RosettaScripts. The XML file is provided below:

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="t14" weights="talaris2014"/>
        <ScoreFunction name="t14w" weights="talaris2014">
                <Reweight scoretype="coordinate_constraint" weight
                    ="1.0"/>
        </ScoreFunction>
    </SCOREFXNS>
    <LIGAND_AREAS>
        <LigandArea name="docking_sidechain" chain="X" cutoff="6.0"
            add_nbr_radius="true" all_atom_mode="true" minimize_ligand
            ="10"/>
        <LigandArea name="final_sidechain" chain="X" cutoff="6.0"
            add_nbr_radius="true" all_atom_mode="true"/>
        <LigandArea name="final_backbone" chain="X" cutoff="7.0"
            add_nbr_radius="false" all_atom_mode="true"
            Calpha_restraints="0.3"/>
    </LIGAND_AREAS>

    <INTERFACE_BUILDERS>
```

```
    <InterfaceBuilder name="side_chain_for_docking" ligand_areas="
        docking_sidechain"/>
    <InterfaceBuilder name="side_chain_for_final" ligand_areas="
        final_sidechain"/>
    <InterfaceBuilder name="backbone" ligand_areas="final_backbone"
        extension_window="3"/>
</INTERFACE_BUILDERS>

<MOVEMAP_BUILDERS>
    <MoveMapBuilder name="docking" sc_interface="
        side_chain_for_docking" minimize_water="true"/>
    <MoveMapBuilder name="final" sc_interface="side_chain_for_final"
         bb_interface="backbone" minimize_water="true"/>
</MOVEMAP_BUILDERS>

<SCORINGGRIDS ligand_chain="X" width="30.0">
    <ClassicGrid grid_name="vdw" weight="1.0" />
</SCORINGGRIDS>

<RESIDUE_SELECTORS>
    <Chain name="ligand" chains="X"/>
    <Not name="receptor" selector="ligand"/>
    <Neighborhood name="interface" selector="ligand" distance
        ="4.0"/>
    <Not name="not_interface" selector="interface"/>
</RESIDUE_SELECTORS>
    <TASKOPERATIONS>
        <RestrictToRepacking name="rtrp"/>
        <InitializeFromCommandline name="ifcl" />
        <OperateOnResidueSubset name="repackonly_ligand" selector="
            ligand" >
                <RestrictToRepackingRLT/>
```

```
        </OperateOnResidueSubset>
        <OperateOnResidueSubset name="repack_interface" selector="
            interface" >
                <RestrictToRepackingRLT/>
        </OperateOnResidueSubset>
        <OperateOnResidueSubset name="fix_notinterface" selector="
            not_interface" >
                <PreventRepackingRLT/>
        </OperateOnResidueSubset>
    </TASKOPERATIONS>
    <SIMPLE_METRICS>
            <TotalEnergyMetric name="energy" />
    </SIMPLE_METRICS>
    <FILTERS>
            <LigInterfaceEnergy name="ifscore"  scorefxn="t14"
                include_cstE="0" energy_cutoff="0.0"/>
            <Rmsd name="rmsd_filter" superimpose="1" threshold="4.0"
                 confidence="1.0"/>
    </FILTERS>
    <SIMPLE_METRICS>
            <InteractionEnergyMetric name="lig_ifscore"
                force_rescore="true" residue_selector="ligand"
                residue_selector2="receptor"  scorefxn="t14" />
    </SIMPLE_METRICS>
<MOVERS>
    # Compute protein-ligand interaction energy
    <RunSimpleMetrics name="ifx" metrics="lig_ifscore" prefix="
        ligand_" />

    # Constraints
    <AddConstraintsToCurrentConformationMover name="cst" coord_dev
        ="1.0" CA_only="0" bb_only="1" bound_width="2.0"/>
```

```
<ClearConstraintsMover name="uncst"/>
<AtomCoordinateCstMover name="relax_cst" coord_dev="1.0" bounded
    ="false" bound_width="0"
sidechain="true" native="false" func_groups="false"/>

# Relax
<VirtualRoot name="root" removable="1" />
<VirtualRoot name="remove" remove="1" />
<FastRelax name="relax"  scorefxn="t14w" ramp_down_constraints="
    false"
repeats="1" task_operations="rtrp,repack_interface,
    fix_notinterface"/>
<ParsedProtocol name="relax_cycle">
        <Add mover_name="root"/>
        <Add mover_name="relax_cst"/>
        <Add mover_name="relax"/>
        <Add mover_name="remove"/>
        <Add mover_name="uncst"/>
</ParsedProtocol>

# Minimize
<MinMover name="min"  scorefxn="t14w" chi="true" bb="true"
    cartesian="false" type="lbfgs_armijo_nonmonotone"/>
<ParsedProtocol name="min_cycle">
        <Add mover_name="cst"/>
        <Add mover_name="min"/>
        <Add mover_name="uncst"/>
</ParsedProtocol>

#### Start BCL Drug Design Movers ####

# Common Type I and II TKI design moves
```

```
# Type I TKI design moves
<BCLFragmentMutateMover name="t1_ewl"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ov_reverse=True,ov_shuffle_h=False,
ring_library=%%rotamer_library%%/ring_libraries/
    drug_ring_database.simple.aro.small.sdf.gz,
extend_within_prob=0.0,direct_link_prob=100000,
druglikeness_type=IsConstitutionDruglike,
mutable_atoms=18)"
/>

<BCLFragmentMutateMover name="t1_ewl_internal"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rotamer_library%%/ring_libraries/
    drug_ring_database.simple.aro.small.sdf.gz,
extend_within_prob=1.0,single_element_link_prob=100000,
N_prob=1000,O_prob=1000,
druglikeness_type=IsConstitutionDruglike,
mutable_atoms=8 10)"
/>

<BCLFragmentMutateMover name="t1_am"
ligand_chain="X"
object_data_label="AddMedChem(
medchem_library=%%medchem_fragments%%/chains.sdf.gz,
druglikeness_type=IsConstitutionDruglike,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1,
fixed_elements=H N O)"
```

```
/>

<BCLFragmentMutateMover name="t1_h_c"
ligand_chain="X"
object_data_label="Halogenate(
druglikeness_type=IsConstitutionDruglike,
mutable_atoms=11 12 13 14 15,
reversible=1,
allowed_halogens=F Cl)"
/>

<BCLFragmentMutateMover name="t1_a_c"
ligand_chain="X"
object_data_label="Alchemy(
druglikeness_type=IsConstitutionDruglike,
mutable_atoms=11 12 13 14 15,
allowed_elements=C,
restrict_to_bonded_h=1)"
/>

<BCLFragmentMutateMover name="t1_rs"
ligand_chain="X"
object_data_label="RingSwap(
ring_library=%%rotamer_library%%/ring_libraries/
    drug_ring_database.simple.aro.sdf.gz,
druglikeness_type=IsConstitutionDruglike,conservative=False,
    restricted=True,
scaffold_mol=%%mutfrag%%,
atom_comparison=ElementType,
bond_comparison=BondOrderOrAromaticWithRingness,
mutable_atoms=11,fix_geometry=True,refine_alignment=False,
ring_initiation_probability=0.0)"
```

```
/>

# Alchemy can be used as a dummy mutate when doing random
    combinations
<BCLFragmentMutateMover name="t1_a_control"
ligand_chain="X"
object_data_label="Alchemy(
druglikeness_type=IsConstitutionDruglike,
mutable_elements=H,
allowed_elements=H
)"
/>

<BCLFragmentMutateMover name="t1_a"
ligand_chain="X"
object_data_label="Alchemy(
druglikeness_type=IsConstitutionDruglike,
mutable_atoms=18,
allowed_elements=N O)"
/>

<RandomMover name="decorate_core" movers="t1_h_c,t1_a_c" weights
    ="0.5,0.5" repeats="2"/>
<RandomMover name="decorate_new" movers="t1_am,t1_a_control"
    weights="0.25,0.75" repeats="1"/>
<RandomMover name="perturb_scaffold" movers="t1_ewl_internal,
    t1_rs,t1_a_control" weights="0.05,0.25,0.7" repeats="1"/>
<RandomMover name="core_mod" movers="perturb_scaffold,
    decorate_core" weights="0.25,0.75" repeats="1"/>

<ParsedProtocol name="design_t1" mode="sequence">
        <Add mover_name="t1_ewl"/>
```

```
        <Add mover_name="decorate_new"/>
        <Add mover_name="core_mod"/>
</ParsedProtocol>

# Type II TKI design moves

<BCLFragmentMutateMover name="t2_ew1"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rotamer_library%%/ring_libraries/individual_rings
    /000.sdf.gz,
extend_within_prob=0.0,amide_link_prob=100000,
    amide_n_attach_prob=0.5,
druglikeness_type=IsConstitutionDruglike,
mutable_atoms=14)"
/>

<BCLFragmentMutateMover name="t2_am"
ligand_chain="X"
object_data_label="AddMedChem(
medchem_library=%%medchem_fragments%%/
    hydrophobic_pocket_fragments.sdf.gz,
druglikeness_type=IsConstitutionDruglike,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1,
fixed_elements=H N O)"
/>

<BCLFragmentMutateMover name="t2_a"
ligand_chain="X"
object_data_label="Alchemy(
druglikeness_type=IsConstitutionDruglike,
```

```
mutable_atoms=18,
allowed_elements=N O)"
/>

<BCLFragmentMutateMover name="add_c"
ligand_chain="X"
object_data_label="Alchemy(
druglikeness_type=IsConstitutionDruglike,
allowed_elements=C,restrict_to_bonded_h=true,
mutable_atoms=14)"
/>

<BCLFragmentMutateMover name="t2_ewl_with_c"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rotamer_library%%/ring_libraries/individual_rings
    /000.sdf.gz,
extend_within_prob=0.0,amide_link_prob=100000,
    amide_n_attach_prob=0.5,
druglikeness_type=IsConstitutionDruglike,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1,
fixed_elements=H N O)"
/>

<ParsedProtocol name="t2_ewl_extended" mode="sequence">
        <Add mover_name="add_c"/>
        <Add mover_name="t2_ewl_with_c"/>
</ParsedProtocol>

<RandomMover name="t2_ewl_rand" movers="t2_ewl,t2_ewl_extended"
    weights="0.9,0.1"/>
```

```
<ParsedProtocol name="design_t2" mode="sequence">
        <Add mover_name="t2_ewl_rand"/>
        <Add mover_name="t2_am"/>
        <Add mover_name="decorate_core"/>
        <Add mover_name="t2_a"/>
</ParsedProtocol>

# Select a design sequence
<RandomMover name="design_t1-2" movers="design_t2,design_t1"
    weights="0.5,0.5"/>

# BCL mutate MCM optimization
<GenericMonteCarlo name="bcl_gmc" mover_name="design_t1-2"
    scorefxn_name="t14w"
trials="20" sample_type="low"
temperature="1.0" drift="0" recover_low="1" reset_baselines="0"
adaptive_movers="0" preapply="0" />

#### End BCL Drug Design Movers ####

# Low resoluteon docking
<Transform name="transform" chain="X" box_size="6.0"
    move_distance="0.2" angle="2.0" cycles="500" repeats="1"
    temperature="%%temp%%" initial_perturb="0.0"
    initial_angle_perturb="0.0" ensemble_proteins="%%confs_list
    %%" use_main_model="false" />

# High resolution docking
<HighResDocker name="high_res_docker" cycles="12"
    repack_every_Nth="3" scorefxn="t14" movemap_builder="docking
    "/>
```

```
        # Scoring
        <InterfaceScoreCalculator name="add_scores" chains="X" scorefxn
            ="t14" compute_grid_scores="0"/>

        <ParsedProtocol name="low_res_dock">
            <Add mover_name="transform"/>
        </ParsedProtocol>

        <ParsedProtocol name="high_res_dock">
                <Add mover_name="relax_cycle"/>
                <Add mover_name="high_res_docker"/>
                <Add mover_name="min_cycle"/>
        </ParsedProtocol>
    </MOVERS>
    <PROTOCOLS>
        <Add mover_name="min_cycle"/>
        <Add mover_name="design_t%%tki_type%%"/>
        <Add mover_name="low_res_dock"/>
        <Add mover_name="high_res_dock"/>
        <Add mover_name="ifx"/>
        <Add mover_name="add_scores"/>
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

The above XML script can be run with the rosetta_scripts application compiled with 'extras=bcl'. Below is a sample Bash script:

```
#!/bin/bash

# Global variables
ROSETTA=Rosetta/main/source/bin/rosetta_scripts.bcl.linuxgccrelease
ROTAMER_LIBRARY=bcl/rotamer_library
MUTABLE_FRAGMENTS=induced_fit_drug_design/lowres/inputs/ligands/LIG.clean.sdf
```

```
MEDCHEM_FRAGMENTS=induced_fit_drug_design/lowres/inputs/medchem_fragments

# Input variables
XML=`readlink -e $1`
PROTEIN=`readlink -e $2`
LIGAND=`readlink -e $3`
PARAMS=$4
TEMP=$5
TYPE=$6
PREFIX=$7

# Derived variables
protein=`basename $PROTEIN .pdb`
ligand=`basename $LIGAND .pdb`

# Run
$ROSETTA \
-parser:protocol $XML \
-in:file:s "$PROTEIN $LIGAND" \
-extra_res_fa "$PARAMS".fa.params \
-extra_res_cen "$PARAMS".cen.params \
-parser:script_vars rotamer_library="${ROTAMER_LIBRARY}" \
-parser:script_vars mutfrag="${MUTABLE_FRAGMENTS}" \
-parser:script_vars medchem_fragments="${MEDCHEM_FRAGMENTS}" \
-parser:script_vars temp="${TEMP}" \
-parser:script_vars tki_type="${TYPE}" \
-parser:script_vars confs_list="${CONFS_LIST}" \
-out:prefix $PREFIX \
-out:pdb_gz true \
-nstruct 100 \
-in:file:fullatom \
-restore_talaris_behavior \
-ignore_zero_occupancy false \
```

```
-linmem_ig 10 \
-mute protocols.qsar.scoring_grid.GridManager \
-constant_seed false > ${PREFIX}.log
```

8.4.3 Induced-fit drug design of PAMs in a mAChR1 cryptic pocket

Design simulations were initiated with the 3-[(1S,2S)-2-hydroxycyclohexyl]-3H,4H-benzo[h]quinazolin-4-one core of BQZ12 excised from the nonplanar "arm" region extending in the cryptic pocket. Chemical perturbations were performed to produce a range of planar (analogous to LY2119620) or nonplanar (analogous to BQZ12) extensions from the core. Following chemical perturbation (design) on the scaffold, Monte Carlo – Metropolis sampling of Y2.64 and C45.50 sidechain rotamers was performed followed by a final minimization of the whole complex.

Once again, we utilized the RosettaScripts framework for the protocol. The following XML file was run with the rosetta_scripts.bcl.linuxgccrelease application.

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="ligand_soft_rep" weights="ligand_soft_rep">
                <Reweight scoretype="fa_elec" weight="0.42"/>
                <Reweight scoretype="hbond_bb_sc" weight="1.3"/>
                <Reweight scoretype="hbond_sc" weight="1.3"/>
                <Reweight scoretype="rama" weight="0.2"/>
        </ScoreFunction>

        <ScoreFunction name="hard_rep" weights="ligand">
                <Reweight scoretype="fa_intra_rep" weight="0.004"/>
                <Reweight scoretype="fa_elec" weight="0.42"/>
                <Reweight scoretype="hbond_bb_sc" weight="1.3"/>
                <Reweight scoretype="hbond_sc" weight="1.3"/>
                <Reweight scoretype="rama" weight="0.2"/>
        </ScoreFunction>
    </SCOREFXNS>
    <LIGAND_AREAS>
        <LigandArea name="docking_sidechain" chain="X" cutoff="6.0"
            add_nbr_radius="true" all_atom_mode="true" minimize_ligand
```

```
        ="10"/>
    <LigandArea name="final_sidechain" chain="X" cutoff="6.0"
        add_nbr_radius="true" all_atom_mode="true"/>
    <LigandArea name="final_backbone" chain="X" cutoff="7.0"
        add_nbr_radius="false" all_atom_mode="true"
        Calpha_restraints="0.3"/>
</LIGAND_AREAS>

<INTERFACE_BUILDERS>
    <InterfaceBuilder name="side_chain_for_docking" ligand_areas="
        docking_sidechain"/>
    <InterfaceBuilder name="side_chain_for_final" ligand_areas="
        final_sidechain"/>
    <InterfaceBuilder name="backbone" ligand_areas="final_backbone"
        extension_window="3"/>
</INTERFACE_BUILDERS>

<MOVEMAP_BUILDERS>
    <MoveMapBuilder name="docking" sc_interface="
        side_chain_for_docking" minimize_water="true"/>
    <MoveMapBuilder name="final" sc_interface="side_chain_for_final"
         bb_interface="backbone" minimize_water="true"/>
</MOVEMAP_BUILDERS>

<RESIDUE_SELECTORS>
    <Chain name="ligand" chains="X"/>
    <Not name="receptor" selector="ligand"/>
    <Index name="gates" resnums="64,244,156" />
    <Neighborhood name="interface" selector="ligand" distance="4"/>
    <Not name="not_interface" selector="interface"/>
    <Not name="not_gates" selector="gates"/>
</RESIDUE_SELECTORS>
```

```
<TASKOPERATIONS>
    <RestrictToRepacking name="rtrp"/>
    <OperateOnResidueSubset name="repack_ligand" selector="ligand" >
            <RestrictToRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name="fix_receptor" selector="receptor"
       >
            <PreventRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name="repack_interface" selector="
       interface" >
            <RestrictToRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name="fix_notinterface" selector="
       not_interface" >
            <PreventRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name="repack_gates" selector="gates" >
            <RestrictToRepackingRLT/>
    </OperateOnResidueSubset>
    <OperateOnResidueSubset name="fix_notgates" selector="not_gates"
        >
            <PreventRepackingRLT/>
    </OperateOnResidueSubset>
</TASKOPERATIONS>

<FILTERS>
    <LigInterfaceEnergy name="ifscore" scorefxn="soft_rep"
       include_cstE="0" energy_cutoff="0.0"/>
    <ScoreType name="score_filter" score_type="total_score"
       threshold="0.0" scorefxn="hard_rep" confidence="1.0"/>
```

```
</FILTERS>
<SIMPLE_METRICS>
    # Score poses
    <InteractionEnergyMetric name="lig_ifscore" force_rescore="1"
        residue_selector="ligand" residue_selector2="receptor"
        scorefxn="hard_rep" />
</SIMPLE_METRICS>
<MOVERS>
    # Constraint movers
    <AddConstraintsToCurrentConformationMover name="cst" coord_dev
        ="1.0" CA_only="0" bb_only="1"/>
    <AtomCoordinateCstMover name="relax_cst" coord_dev="1.0" bounded
        ="false" bound_width="0" sidechain="true" native="false"
        func_groups="false"/>
    <ClearConstraintsMover name="uncst"/>

    # Minimization movers
    <MinMover name="min_soft" scorefxn="ligand_soft_rep" chi="true"
        bb="true" cartesian="false" type="lbfgs_armijo_nonmonotone
        "/>
    <MinMover name="min_hard" scorefxn="hard_rep" chi="true" bb="
        true" cartesian="false" type="lbfgs_armijo_nonmonotone"/>
    <ParsedProtocol name="min_cycle_soft">
            <Add mover_name="cst"/>
            <Add mover_name="min_soft"/>
            <Add mover_name="uncst"/>
    </ParsedProtocol>
    <ParsedProtocol name="min_cycle_hard">
            <Add mover_name="cst"/>
            <Add mover_name="min_hard"/>
            <Add mover_name="uncst"/>
    </ParsedProtocol>
```

```
# Relax
<VirtualRoot name="root" removable="1" />
<VirtualRoot name="remove" remove="1" />
<FastRelax name="relax" scorefxn="hard_rep"
    ramp_down_constraints="false" repeats="2" task_operations="
    rtrp,repack_interface,fix_notinterface"/>
<ParsedProtocol name="relax_cycle">
        <Add mover_name="root"/>
        <Add mover_name="relax_cst"/>
        <Add mover_name="relax"/>
        <Add mover_name="remove"/>
        <Add mover_name="uncst"/>
</ParsedProtocol>

# Packing
<PackRotamersMover name="pack" scorefxn="hard_rep" nloop="50"
    task_operations="rtrp,repack_ligand,fix_receptor"/>

# Backbone sampling (coupled with sidechains)
<Backrub name="backrub"/>
<Shear name="shear" residue_selector="interface" scorefxn="
    soft_rep" temperature="0.593" preserve_detailed_balance
    ="1"/>
Sidechain name="sidechain" preserve_detailed_balance="1"
    task_operations="rtrp,repack_interface,fix_notinterface"
    prob_withinrot="0.5"/>
<Sidechain name="sidechain" preserve_detailed_balance="1"
    task_operations="rtrp,repack_gates,fix_notgates"
    prob_withinrot="0.5"/>
<ParsedProtocol name="pseudo_coupled_moves">
        Add mover_name="backrub"/>
```

```
        Add mover_name="shear"/>
        <Add mover_name="sidechain"/>
</ParsedProtocol>
<GenericMonteCarlo name="run_pcm" mover_name="
    pseudo_coupled_moves"
trials="1000" sample_type="low" filter_name="score_filter"
temperature="0.593" drift="1" recover_low="1" reset_baselines
    ="0"
adaptive_movers="0" preapply="0" />

# High resolution perturbation; default values: cycles=6,
    repack_every_Nth=3
<HighResDocker name="high_res_docker" cycles="6"
    repack_every_Nth="3" scorefxn="ligand_soft_rep"
    movemap_builder="docking"/>
<FinalMinimizer name="final" scorefxn="hard_rep" movemap_builder
    ="final"/>

# Scoring
<InterfaceScoreCalculator name="interaction_energy" chains="X"
    scorefxn="hard_rep" compute_grid_scores="0"/>

# Single design moves
<BCLFragmentMutateMover name="empty" ligand_chain="X"
    object_data_label="Alchemy(allowed_elements=H,
    restrict_to_bonded_h=true)" />

<BCLFragmentMutateMover name="add_c"
ligand_chain="X"
object_data_label="Alchemy(
druglikeness_type=None,
allowed_elements=C,restrict_to_bonded_h=true,
```

```
mutable_atoms=9)"
/>

<BCLFragmentMutateMover name="amide_link_aro_ring"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rings%%,
extend_within_prob=0.0,amide_link_prob=100000,
    amide_n_attach_prob=0.5,
druglikeness_type=None,
mutable_atoms=9)"
/>

<BCLFragmentMutateMover name="single_ele_link_aro_ring"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rings%%,
extend_within_prob=0.0,
single_element_link_prob=100000,
O_prob=10000, S_prob=10000, N_prob=10000,
druglikeness_type=None,
mutable_atoms=9)"
/>

<BCLFragmentMutateMover name="direct_link_aro_ring"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rings%%,
extend_within_prob=0.0,
direct_link_prob=100000,
druglikeness_type=None,
mutable_atoms=9)"
```

```
/>

<BCLFragmentMutateMover name="add_medchem"
ligand_chain="X"
object_data_label="AddMedChem(
medchem_library=%%fragments%%,
druglikeness_type=None,
mutable_atoms=9)"
/>

<BCLFragmentMutateMover name="amide_link_aro_ring_basefrag"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rings%%,
extend_within_prob=0.0,amide_link_prob=100000,
    amide_n_attach_prob=0.5,
druglikeness_type=None,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1,
fixed_elements=H C)"
/>

<BCLFragmentMutateMover name="direct_link_aro_ring_basefrag"
ligand_chain="X"
object_data_label="ExtendWithLinker(
ring_library=%%rings%%,
extend_within_prob=0.0,direct_link_prob=100000,
druglikeness_type=None,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1,
fixed_elements=H N O)"
/>
```

```
<BCLFragmentMutateMover name="add_medchem_basefrag"
ligand_chain="X"
object_data_label="AddMedChem(
medchem_library=%%fragments%%,
druglikeness_type=None,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1,
fixed_elements=H N O)"
/>

<BCLFragmentMutateMover name="halogenate_basefrag"
ligand_chain="X"
object_data_label="Halogenate(
druglikeness_type=None,
allowed_halogens=F Cl,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1)"
/>

<BCLFragmentMutateMover name="alchemy_basefrag"
ligand_chain="X"
object_data_label="Alchemy(
druglikeness_type=None,
allowed_elements=C,
restrict_to_bonded_h=1,
mutable_fragments=%%mutfrag%%,
complement_mutable_fragments=1,
fixed_elements=H O S)"
/>

# Combo design moves
```

```
<RandomMover name="planar_link"
movers="direct_link_aro_ring,amide_link_aro_ring"
weights="0.5,0.5"
/>

<RandomMover name="planar_link_basefrag"
movers="direct_link_aro_ring_basefrag,
    amide_link_aro_ring_basefrag"
weights="0.5,0.5"
/>

<ParsedProtocol name="c_link_aro_ring" mode="sequence">
        <Add mover_name="add_c"/>
        <Add mover_name="planar_link_basefrag"/>
</ParsedProtocol>

<RandomMover name="nonplanar_link"
movers="single_ele_link_aro_ring,c_link_aro_ring"
weights="0.5,0.5"
/>

<RandomMover name="decorate_big"
movers="direct_link_aro_ring_basefrag,add_medchem_basefrag,empty
    "
weights="0.33,0.33,0.34"
/>

<RandomMover name="decorate_small"
movers="halogenate_basefrag,alchemy_basefrag,empty"
weights="0.33,0.33,0.34"
/>
```

```
<RandomMover name="decorate"
movers="halogenate_basefrag,add_medchem_basefrag,empty"
weights="0.33,0.33,0.34"
/>

<RandomMover name="decorate_halogens"
movers="halogenate_basefrag,empty"
weights="0.5,0.5"
/>

<ParsedProtocol name="run_a" mode="sequence">
        <Add mover_name="planar_link"/>
        <Add mover_name="decorate_big"/>
        <Add mover_name="decorate_small"/>
</ParsedProtocol>

<ParsedProtocol name="run_b" mode="sequence">
        <Add mover_name="nonplanar_link"/>
        <Add mover_name="decorate_big"/>
        <Add mover_name="decorate_small"/>
</ParsedProtocol>

<ParsedProtocol name="run_c" mode="sequence">
        <Add mover_name="nonplanar_link"/>
        <Add mover_name="direct_link_aro_ring_basefrag"/>
        <Add mover_name="decorate"/>
</ParsedProtocol>

<ParsedProtocol name="run_d" mode="sequence">
        <Add mover_name="single_ele_link_aro_ring"/>
        <Add mover_name="direct_link_aro_ring_basefrag"/>
        <Add mover_name="decorate_halogens"/>
```

```
        </ParsedProtocol>

        <RandomMover name="run" movers="run_a,run_d" weights="0.2,0.8"/>
    </MOVERS>

    <PROTOCOLS>
        <Add mover_name="run"/>
        <Add mover_name="pack"/>
        <Add mover_name="min_cycle_soft"/>
        <Add mover_name="run_pcm"/>
        <Add mover_name="final"/>
        <Add mover_name="interaction_energy"/>
        <Add metrics="lig_ifscore"/>
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

We used the following Bash script to run the RosettaScripts application:

```
#!/bin/bash
# Global variables
ROSETTA=Rosetta/main/source/bin/rosetta_scripts.bcl.linuxgccrelease

# Input variables
XML=`readlink -e $1`
PROTEIN=`readlink -e $2`
LIGAND=`readlink -e $3`
PARAMS=`readlink -e $4`
PREFIX=$5
RINGS=bcl/rotamer_library/ring_libraries/drug_ring_database.simple.aro.
    small.sdf.gz
FRAGMENTS=bcl/rotamer_library/medchem_fragments/bcl_buildfrag_0.sdf.gz
MUTFRAG=BQ0.clean.sdf

# Run
```

```
$ROSETTA \
-parser:protocol $XML \
-in:file:s "$PROTEIN $LIGAND" \
-parser:script_vars prefix="${PREFIX}" \
-parser:script_vars rings=${RINGS} \
-parser:script_vars fragments=${FRAGMENTS} \
-parser:script_vars mutfrag=${MUTFRAG} \
-parser:script_vars progress_file="${PREFIX}.gmc.log" \
-extra_res_fa ${PARAMS} \
-out:prefix $PREFIX \
-out:pdb_gz true \
-packing:ex1 true \
-packing:ex2 true \
-nstruct 100 \
-in:file:fullatom \
-restore_pre_talaris_2013_behavior \
-score:weights ligand \
-ignore_zero_occupancy false \
-mute protocols.rosetta_scripts.ParsedProtocol protocols.monte_carlo.
    GenericMonteCarloMover \
-overwrite \
-linmem_ig 10 #> ${PREFIX}.log
```

CHAPTER 9

Conclusions and future directions

9.1 Summary and Implications

Cheminformatics and computer-aided drug design (CADD) have matured substantially in the last decade. CADD is no longer the niche approach of a subset of specialists, but rather an integral component of the drug discovery process in both academia and industry (Macalino et al., 2015). As adoption of CADD continues to spread, so too will the demand for increasingly robust methods that leave no target undruggable. Today, we strive for precision in drug design, such that our molecules bind to specific conformations of flexible proteins and are selective against homologous receptors and/or mutants.

The need for such precision in drug design is apparent when we consider epidermal growth factor receptor (EGFR)-mutant non-small cell lung cancer (NSCLC). Proteins can be conformationally dynamic biomolecules. These dynamics give rise to function, and aberrant dynamics can lead to disease. Changes in protein dynamics of EGFR caused by amino acid mutations drive oncogenic behavior in NSCLC (Brown et al., 2019a; Du et al., 2021; Hanker et al., 2021; Shan et al., 2013, 2012). In Chapters 2-5, we characterized the mechanisms of oncogenesis and tyrosine kinase inhibitor (TKI) resistance in several new EGFR, HER2, and HER3 variants, as well as identified structure-function relations that may be responsible for variable outcomes in NSCLC patients with different EGFR exon 19 deletion variants (Brown et al., 2019a; Du et al., 2021; Hanker et al., 2021).

It is crucial that CADD methods continue to improve to enable precision targeting of EGFR and other receptors. In Chapters 6-8, we describe new cheminformatics tools that collectively enable a customizable framework for small molecule drug design (Brown et al., 2019b, 2021, 2022). This framework leverages the combined abilities of the BioChemical Library (BCL) and Rosetta to offer unique advantages.

The BCL code is utilized to enable multiple routes for chemical perturbation during design. This means that users can create new molecules using explicitly defined chemical reactions, or they can employ medicinal chemistry-inspired alchemical transformations. These perturbations use internal graph- and property-based molecular alignment schemes to minimize real space perturbation of non-mutable components of the molecular scaffolds. The BCL descriptor framework can be used to constrain the drug design space to druglike chemical space. The same descriptor framework gives access to machine learning (ML) score functions, such as the decomposable BCL-AffinityNet score described in Chapter 7.

The Rosetta code enables conformational sampling of the target receptor during design. This allows

induced-fit and conformational selection to be directly accounted for during design using any combination of kinematic functions available in Rosetta. Rosetta also maintains an arsenal of tools for protein design. Some drug design simulations, such as for designer receptors exclusively activated by designer drugs (DREADDs) and other chemogenetics tasks, were previously beyond the scope of CADD.

The new capabilities for drug design made available through the work presented in Chapters 6-8 are poised to accelerate drug discovery for challenges identified in Chapters 2-5.

9.2 Limitations and Future Directions

9.2.1 On the use of biased sampling approaches to model hyperstable kinase variants

A recurring theme in chapters 2 - 5 was the use of molecular dynamics (MD) simulations to map changes to the conformational free energy landscape (FEL) of oncogenic kinase mutants. The FEL in these studies is a conformational pathway between two or more states that have biological significance. Here, we are frequently interested in mapping the energetic differences in the the active and inactive states of the kinase domain because oncogenic variants are frequently pathogenic by virtue of their excess signalling.

Defining the boundaries between the active and inactive states is typically readily achievable; extensive crystallographic studies over the last two decades have provided a wealth of information on the activation states of kinases. The challenge is in connecting them via a biologically relevant conformational pathway. Frequently, we are unable to naively sample state transitions, such as in Chapter 3 with the oncogenic ex19del variants.

In our studies, we combined steered MD (SMD) and umbrella sampling (US) simulations to first generate a low dimensional transition path and then sample distributions along along discrete intervals of the path, respectively. The benefit of this approach is that the relative probabilities of each discrete interval along the chosen low dimensional collective variable(s) can be estimated, which enables the construction of a relative free energy surface in that space. Using this approach, one can obtain differences in the thermodynamic stability of different states, identify energetic minima and maxima, and estimate transition rates between states. A limitation of the SMD+US approach is that the generated pathway may not be biologically relevant, even if the end-states are experimentally validated. Thus, even if we determine the correct free energy difference between end-states, alternately identified minima and maxima as well as kinetic information can be inaccurate.

An alternative strategy would be to use MD simulations to equilibrate ensembles of the distinct functionally relevant states and then perform $\Delta\Delta G$ calculations on samples from each state using Rosetta. The Rosetta energy function contains a mix of potential energy terms and statistically-derived terms, such that the Rosetta energy in principle can be likened to a folding free energy. Thus, taking the average Rosetta

energy across an ensemble of MD simulations conformers belonging to a single state (e.g., the active state) provides an estimate of the ΔG_{fold} of that state in Rosetta energy units (REU). The difference in $\Delta G_{active,fold}$ and $\Delta G_{inactive,fold}$ estimates of the relative stability of the active and inactive states. However, there are at least two primary limitations to this approach. First, the physical meaning of REUs and their differences are less clear than standard units such as kcal/mol, which make it difficult to interpret results in the absence of a high-quality calibration curve. Second, such an approach explicitly neglects a transition path, and thus provides no information on kinetic barriers of biological significance. Despite these limitations, this approach requires fewer computational resources and warrants additional investigation.

9.2.2 Toward novel therapeutic modalities in precision oncology

On the one hand, our studies suggest that the conformational propensities of oncogenic variants can potentially be leveraged to develop mutant-selective small molecule inhibitors. In our work, this is perhaps most evident through EGFR E746_S752>V/G724S. This mutant escapes inhibition by the current first-line therapy osimertinib, a third-generation covalent TKI, by stabilizing a rare conformation in which the glycine-rich loop F723 is unable to form a favorable stacking interaction with the indole ring of osimertinib (Chapter 2). One can imagine that this unique conformational state can be directly targeted with structure-based drug design to create a derivative of an existing FDA-approved TKI.

But on the other hand, we may be reaching a plateau in efficacy of single-agent TKIs as therapeutic agents in cancer. This plateau is not driven by certain mutants being undruggable. Quite the opposite - a major barrier to TKI efficacy in clinical practice is toxicity and side effect tolerability from off-target (e.g., wild-type) inhibition (Lin et al., 2019). We show that the sensitivity of common oncogenic EGFR variants to ATP-competitive TKIs relative to wild-type EGFR is the result of reduced ATP binding affinity. Not only can resistance mutations revert this loss of ATP binding affinity to reduce TKI efficacy, we demonstrate that oncogenic variants themselves may be less amenable to targeting because of higher ATP binding affinities. This results in the need to make higher affinity TKIs, which have a higher likelihood of also inhibiting wild-type EGFR and leading to intolerable side effects.

Perhaps ongoing improvement in the accuracy and efficiency of binding energy calculations in CADD software packages will alleviate this concern; however, energy function improvement is not sufficient. We will need to explicitly model multiple equilibria during the drug design process to optimize selectivity for mutant, selectivity for conformation, and avidity against native substrates.

An alternative approach would be to preemptively determine the residues that are capable of mutating into resistance mutations prior to designing a drug. This would potentially allow drug developers to optimize interactions with functionally conserved residues less prone to on-target resistance.

Still another approach would be to move beyond single-agent TKIs altogether. There are a variety of emerging therapeutics, such as antibody-drug conjugates (ADCs) and proteolysis targeting chimeras (PROTACs), that strategically combine small molecules and proteins to achieve the desired therapeutic effect. We anticipate that the BCL-Rosetta integration tools for drug design described in Chapter 8 will be valuable as we seek to expand CADD methods in the direction of these new modalities.

9.2.3 Next steps in the development of BCL drug design chemical space perturbations

We developed a series of drug design chemical perturbations, or "mutates", that are capable of performing one-shot made-on-demand style reactions, single- and multi-component reactions, or reaction-free medicinal chemistry-inspired "alchemical" perturbations in the BCL software package. These mutates enable a wide variety of chemical alterations to be made to chemical scaffolds; however, one can imagine many more complex design operations that are inaccessible with the current implementations.

Macrocyclic closure of a single fragment to reduce flexibility, for example, cannot currently be performed. Macrocycle formation is an attractive strategy to reduce conformational degrees of freedom, and thus entropic cost of binding, to otherwise promising molecules. We also currently can only perform 2-component using the convenient made-on-demand style reaction format; 3- and 4-component reactions require use of the MDL RXN file format. Another common design operation that we currently cannot perform with the BCL drug design framework is fragment linking. Currently there is no mutate to generate linkers between two distinct chemical fragments. This is a critical design task in fragment-based drug discovery and PROTAC design.

Fortunately, the drug design framework is extensible and can readily accommodate additional algorithms. For example, one potential approach to add a linker design algorithm would be to create a class that derives from the FragmentMutateInterface base class and utilizes the ExtendWithLinker perturbation to build linkers with customizable compositions. Given two disconnected fragments and an attachment-site atom on each fragment, extend one-half of the full linker from each attachment site. Rapidly generate conformational ensembles of just the half-linker regions of the two fragments without perturbing the starting fragment and filter for compatible pairs by the pairwise distance between the terminal atoms of the half-linkers (discarding pairs that are unlikely to result in successful linker closures). For all compatible pairs, join the half-linkers to yield the fully-connected linker. Perform local conformational sampling of the linker and select the final 3D conformer as the one with the best BCL ConfScore that has an RMSD to the starting fragment within some tolerance.

One limitation of the above-described algorithm is the dependence on sampling conformers in torsion-space to refine 3D structures. The conformational ensembles of the half-linkers would not result in any perturbation to the starting fragments, but the local ensembles of the linker in the fully-connected molecule

can cause lever-arm effects that alter the starting fragment Cartesian-space positions. Moreover, it adds a sampling overhead to perform a simple refinement, which reduces the efficiency of the algorithm. It would be more efficient, and likely more effective, to perform gradient-based minimization of the linker with restraints on the positions of the starting fragments atoms for the final refinement step. Thus, another important enhancement to the BCL drug design framework will be the addition of a small molecule molecular mechanics force field for such calculations.

9.2.4 Incorporating optimization tools into the BCL drug design framework

The current philosophy in the BCL drug design framework is to enable user control over the design process as much as possible - from the selection of fixed and mutable atom indices, elements, and fragments, to the selection of the nature of the chemical perturbations and composition of external chemical libraries used in design. If the user provides very specific instructions on how a mutate or series of mutates are intended to proceed, the outcome of the design process can be completely determined at runtime. If the user input is less explicit, however, there are several stochastic components that can lead to different outcomes.

Indeed, the default behavior of the alchemical framework in the absence of user restrictions is to apply the selected mutate to a randomly chosen atom. Many of the mutates, such as Alchemy, ExtendWithLinker, RingSwap, etc., are also stochastic with respect to the chemical perturbation they apply. The default behavior of one-shot reaction-based design approach is to randomly combine complementary fragments from an external library at the prescribed attachment sites. Finally, the default behavior of the MDL RXN reaction design framework is to randomly identify a reaction into which the starting fragment can be a reagent, identify suitable co-reagents, and apply the random reaction.

Clearly, there is a lot of stochasticity. This is because the current drug design framework is setup to enable sampling control at the user level. When the framework was originally designed, it was intended that optimization routines would run external to the design framework, such that the distribution from which the optimizer samples is a random distribution (limited by any user restrictions). Examples of such optimization routines include Monte Carlo - Metropolis sampling and evolutionary algorithms.

However, an additional approach would be to enable some level of optimization within the drug design framework to increase the efficiency of external optimizers. This could allow mutate-specific parameter tuning to occur prior to generation of the final ensemble of candidate molecules. Consider the alchemical drug design framework. A simple example of this would be in the selection of mutable atom indices for each in a set of mutates. Another example would be the sequence in which the mutates are applied to the starting scaffold. Alternatively, consider either the one-shot or MDL RXN reaction design approaches. The parameters dictating the probabilities of choosing specific reactions and/or reagents can be pre-tuned prior to

the end-stage design simulation. There is substantial room to expand the capabilities of the BCL drug design framework.

9.2.5 Expanding the BCL-Rosetta integration to enable polymeric design with exotic chemical modifications

The BCL-Rosetta integration enables on-the-fly structure-based all-atom design of small molecules in a flexible binding pocket. Currently, this functionality does not include atom-based design of polymers, such as peptides, proteins, and nucleic acids. Polymer design in Rosetta still occurs at the residue-level using pre-generated parameters files for non-standard residue types.

The next step in the BCL-Rosetta integration is enable polymer design at the atom level as an alternative strategy to residue-based design. In addition to expanding the types of algorithms that can be developed to optimize polymer sequences, atom-based approaches require potentially substantially less user preparation - you may generate e.g., 10,000 unique phenylalanine derivatives without have to design structure files and parameters for the derivatives prior to running the simulation.

9.2.6 On constructing modular interfaces in Rosetta to maximize out-of-the-box integration with the BCL

We integrated BCL into the Rosetta codebase as an external submodule enabled at compile-time. This means that all BCL data structures are available at the C++ level in Rosetta for any developer to access. It also means that the BCL can be developed actively in tandem with Rosetta, requiring only that the submodule being referenced by Rosetta at compile-time is updated to the BCL developer code.

At the user level, however, currently the BCL-Rosetta integration is only relevant if the user is performing drug design with the new BCLFragmentMutateMover in Rosetta. An ongoing goal is to identify BCL code of broad interest to the Rosetta community and create interfaces for them in higher level Rosetta API, such as RosettaScripts XML and PyRosetta. For example, exposing the BCL descriptor framework to Rosetta SimpleMetrics or Filters would be broadly applicable to small molecule, peptide, and potentially protein design projects, especially those involving non-canonical amino acids and other exotic residues.

9.2.7 Interoperability of functionally orthogonal software packages for drug discovery and design

Beyond the BCL-Rosetta integration, there are a number of exceptional software packages for drug discovery that have unique strengths. Indeed, there is RosettaCommons community support behind developing protocols to incorporate deep learning models and packages into Rosetta to facilitate complex protocol development for structure prediction and design. Similarly, creating cross-talk between Rosetta and molecular

mechanics packages would enable more seamless protocol development. For example, creating classes and utilities to more easily interconvert Rosetta, BCL, OpenMM, and Parmed data structures would facilitate protocol development that combines chemogenetics drug design (BCL-Rosetta) with free energy perturbation for binding affinity prediction (OpenMM-Parmed). There is currently an unprecedented opportunity for cooperation and synergism in CADD, and I believe that by building on these opportunities we will be able to provide effective therapeutics for previously untreatable diseases.

References

Paul Labute, Chris Williams, Miklos Feher, Elizabeth Sourial, and Jonathan M. Schmidt. Flexible alignment of small molecules. *Journal of Medicinal Chemistry*, 44(10):1483–1490, 2001. ISSN 0022-2623. doi: 10.1021/jm0002634. URL http://dx.doi.org/10.1021/jm0002634http://pubs.acs.org/doi/full/10.1021/jm0002634.

Shek Ling Chan and Paul Labute. Training a scoring function for the alignment of small molecules. *Journal of Chemical Information and Modeling*, 50(9):1724–1735, 2010. ISSN 1549-9596. doi: 10.1021/ci100227h. URL http://dx.doi.org/10.1021/ci100227hhttp://pubs.acs.org/doi/full/10.1021/ci100227h.

Dongsheng Zhu, Huocong Huang, Daniel M. Pinkas, Jinfeng Luo, Debolina Ganguly, Alice E. Fox, Emily Arner, Qiuping Xiang, Zheng-Chao Tu, Alex N. Bullock, Rolf A. Brekken, Ke Ding, and Xiaoyun Lu. 2-amino-2,3-dihydro-1h-indene-5-carboxamide-based discoidin domain receptor 1 (ddr1) inhibitors: Design, synthesis, and in vivo antipancreatic cancer efficacy. *Journal of Medicinal Chemistry*, 62(16):7431–7444, 2019. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.9b00365. URL https://doi.org/10.1021/acs.jmedchem.9b00365.

Stephani Joy Y. Macalino, Vijayakumar Gosu, Sunhye Hong, and Sun Choi. Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*, 38(9):1686–1701, 2015. ISSN 1976-3786. doi: 10.1007/s12272-015-0640-5. URL https://doi.org/10.1007/s12272-015-0640-5.

G. Sliwoski, S. Kothiwale, J. Meiler, and Jr. Lowe, E. W. Computational methods in drug discovery. *Pharmacol Rev*, 66(1):334–95, 2014. ISSN 1521-0081 (Electronic) 0031-6997 (Linking). doi: 10.1124/pr.112.007336. URL http://www.ncbi.nlm.nih.gov/pubmed/24381236http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3880464/pdf/pr.112.007336.pdf.

Xin Yang, Yifei Wang, Ryan Byrne, Gisbert Schneider, and Shengyong Yang. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*, 119(18):10520–10594, 2019. ISSN 0009-2665. doi: 10.1021/acs.chemrev.8b00728. URL https://doi.org/10.1021/acs.chemrev.8b00728. doi: 10.1021/acs.chemrev.8b00728.

Sumudu P. Leelananda and Steffen Lindert. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry*, 12:2694–2718, 2016. ISSN 1860-5397. doi: 10.3762/bjoc.12.267. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5238551/.

A. R. Geanes, H. P. Cho, K. D. Nance, K. M. McGowan, P. J. Conn, C. K. Jones, J. Meiler, and C. W. Lindsley. Ligand-based virtual screen for the discovery of novel m5 inhibitor chemotypes. *Bioorg Med Chem Lett*, 26(18):4487–91, 2016. ISSN 1464-3405 (Electronic) 0960-894X (Linking). doi: 10.1016/j.bmcl.2016.07.071. URL https://www.ncbi.nlm.nih.gov/pubmed/27503678http://ac.els-cdn.com/S0960894X16307995/1-s2.0-S0960894X16307995-main.pdf?_tid=3d36eaa8-a8ef-11e6-9c5a-00000aab0f01&acdnat=1478965796_25051127273012ff706fade68825f619.

M. Butkiewicz, Jr. Lowe, E. W., R. Mueller, J. L. Mendenhall, P. L. Teixeira, C. D. Weaver, and J. Meiler. Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules*, 18(1):735–56, 2013. ISSN 1420-3049 (Electronic) 1420-3049 (Linking). doi: 10.3390/molecules18010735. URL http://www.ncbi.nlm.nih.gov/pubmed/23299552.

Reed M. Stein, Hye Jin Kang, John D. McCorvy, Grant C. Glatfelter, Anthony J. Jones, Tao Che, Samuel Slocum, Xi-Ping Huang, Olena Savych, Yurii S. Moroz, Benjamin Stauch, Linda C. Johansson, Vadim Cherezov, Terry Kenakin, John J. Irwin, Brian K. Shoichet, Bryan L. Roth, and Margarita L. Dubocovich. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature*, 579(7800):609–614, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2027-0. URL https://doi.org/10.1038/s41586-020-2027-0.

Lingle Wang, Jennifer Chambers, and Robert Abel. *Protein–Ligand Binding Free Energy Calculations with FEP+*, pages 201–232. Springer New York, New York, NY, 2019a. ISBN 978-1-4939-9608-7. doi: 10.1007/978-1-4939-9608-7_9. URL https://doi.org/10.1007/978-1-4939-9608-7_9.

Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K. Dahlgren, Jeremy Greenwood, Donna L. Romero, Craig Masse, Jennifer L. Knight, Thomas Steinbrecher, Thijs Beuming, Wolfgang Damm, Ed Harder, Woody Sherman, Mark Brewer, Ron Wester, Mark Murcko, Leah Frye, Ramy Farid, Teng Lin, David L. Mobley, William L. Jorgensen, Bruce J. Berne, Richard A. Friesner, and Robert Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015. ISSN 0002-7863. doi: 10.1021/ja512751q. URL https://doi.org/10.1021/ja512751q.

Junjie Zou, Chuan Tian, and Carlos Simmerling. Blinded prediction of protein-ligand binding affinity using amber thermodynamic integration for the 2018 d3r grand challenge 4. *Journal of computer-aided molecular design*, 33(12):1021–1029, 2019. ISSN 1573-4951 0920-654X. doi: 10.1007/s10822-019-00223-x. URL https://pubmed.ncbi.nlm.nih.gov/31555923https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6899192/. 31555923[pmid] PMC6899192[pmcid] 10.1007/s10822-019-00223-x[PII].

William L. Jorgensen and Laura L. Thomas. Perspective on free-energy perturbation calculations for chemical equilibria. *Journal of Chemical Theory and Computation*, 4(6):869–876, 2008. ISSN 1549-9618. doi: 10.1021/ct800011m. URL https://doi.org/10.1021/ct800011m. doi: 10.1021/ct800011m.

Jianing Lu, Xuben Hou, Cheng Wang, and Yingkai Zhang. Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. *Journal of Chemical Information and Modeling*, 59(11):4540–4549, 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00645. URL https://doi.org/10.1021/acs.jcim.9b00645.

Benjamin P. Brown, Jeffrey Mendenhall, Alexander R. Geanes, and Jens Meiler. General purpose structure-based drug discovery neural network score functions with human-interpretable pharmacophore maps. *Journal of Chemical Information and Modeling*, 61(2):603–620, 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c01001. URL https://doi.org/10.1021/acs.jcim.0c01001. doi: 10.1021/acs.jcim.0c01001.

James Kirkpatrick, Brendan McMorrow, H. P. Turban David, L. Gaunt Alexander, S. Spencer James, G. D. G. Matthews Alexander, Annette Obika, Louis Thiry, Meire Fortunato, David Pfau, Román Castellanos Lara, Stig Petersen, W. R. Nelson Alexander, Pushmeet Kohli, Paula Mori-Sánchez, Demis Hassabis, and J. Cohen Aron. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science*, 374(6573):1385–1389, 2021. doi: 10.1126/science.abj6511. URL https://doi.org/10.1126/science.abj6511. doi: 10.1126/science.abj6511.

Francesco Gentile, Jean Charle Yaacoub, James Gleave, Michael Fernandez, Anh-Tien Ton, Fuqiang Ban, Abraham Stern, and Artem Cherkasov. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nature Protocols*, 17(3):672–697, 2022. ISSN 1750-2799. doi: 10.1038/s41596-021-00659-2. URL https://doi.org/10.1038/s41596-021-00659-2.

Francesco Gentile, Vibudh Agrawal, Michael Hsing, Anh-Tien Ton, Fuqiang Ban, Ulf Norinder, Martin E. Gleave, and Artem Cherkasov. Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, 6(6):939–949, 2020. ISSN 2374-7943. doi: 10.1021/acscentsci.0c00229. URL https://doi.org/10.1021/acscentsci.0c00229. doi: 10.1021/acscentsci.0c00229.

Louis Bellmann, Patrick Penner, Marcus Gastreich, and Matthias Rarey. Comparison of combinatorial fragment spaces and its application to ultralarge make-on-demand compound catalogs. *Journal of Chemical Information and Modeling*, 62(3):553–566, 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c01378. URL https://doi.org/10.1021/acs.jcim.1c01378. doi: 10.1021/acs.jcim.1c01378.

Robert Schmidt, Raphael Klein, and Matthias Rarey. Maximum common substructure searching in combinatorial make-on-demand compound spaces. *Journal of Chemical Information and Modeling*, 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00640. URL https://doi.org/10.1021/acs.jcim.1c00640. doi: 10.1021/acs.jcim.1c00640.

Arman A. Sadybekov, Anastasiia V. Sadybekov, Yongfeng Liu, Christos Iliopoulos-Tsoutsouvas, Xi-Ping Huang, Julie Pickett, Blake Houser, Nilkanth Patel, Ngan K. Tran, Fei Tong, Nikolai Zvonok, Manish K. Jain, Olena Savych, Dmytro S. Radchenko, Spyros P. Nikas, Nicos A. Petasis, Yurii S. Moroz, Bryan L. Roth, Alexandros Makriyannis, and Vsevolod Katritch. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature*, 601(7893):452–459, 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04220-9. URL https://doi.org/10.1038/s41586-021-04220-9.

Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7):eaap7885. doi: 10.1126/sciadv.aap7885. URL https://doi.org/10.1126/sciadv.aap7885. doi: 10.1126/sciadv.aap7885.

Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy, Maksim D. Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R. Shayakhmetov, Alexander Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy, Lennart H. Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature Biotechnology*, 37 (9):1038–1040, 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0224-x. URL https://doi.org/10.1038/s41587-019-0224-x.

Benjamin P. Brown, Oanh Vu, Alexander R. Geanes, Sandeepkumar Kothiwale, Mariusz Butkiewicz, Jr. Lowe, Edward W., Ralf Mueller, Richard Pape, Jeffrey Mendenhall, and Jens Meiler. Introduction to the biochemical library (bcl): An application-based open-source toolkit for integrated cheminformatics and machine learning in computer-aided drug discovery. *Frontiers in pharmacology*, 13: 833099–833099, 2022. ISSN 1663-9812. doi: 10.3389/fphar.2022.833099. URL https://pubmed.ncbi.nlm.nih.gov/35264967https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8899505/. 35264967[pmid] PMC8899505[pmcid] 833099[PII].

Samuel Boobier, David R. J. Hose, A. John Blacker, and Bao N. Nguyen. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications*, 11(1):5753, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19594-z. URL https://doi.org/10.1038/s41467-020-19594-z.

Lucas Moreno and Andrew D. J. Pearson. How can attrition rates be reduced in cancer drug discovery? *Expert Opinion on Drug Discovery*, 8(4):363–368, 2013. ISSN 1746-0441. doi: 10.1517/17460441.2013.768984. URL https://doi.org/10.1517/17460441.2013.768984. doi: 10.1517/17460441.2013.768984.

Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace, and Alex Weir. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, 14(7):475–486, 2015. ISSN 1474-1784. doi: 10.1038/nrd4609. URL https://doi.org/10.1038/nrd4609.

Richard K. Harrison. Phase ii and phase iii failures: 2013–2015. *Nature Reviews Drug Discovery*, 15(12): 817–818, 2016. ISSN 1474-1784. doi: 10.1038/nrd.2016.184. URL https://doi.org/10.1038/nrd.2016.184.

Susan Klaeger, Stephanie Heinzlmeir, Mathias Wilhelm, Harald Polzer, Binje Vick, Paul-Albert Koenig, Maria Reinecke, Benjamin Ruprecht, Svenja Petzoldt, Chen Meng, Jana Zecha, Katrin Reiter, Huichao Qiao, Dominic Helm, Heiner Koch, Melanie Schoof, Giulia Canevari, Elena Casale, Stefania Re Depaolini, Annette Feuchtinger, Zhixiang Wu, Tobias Schmidt, Lars Rueckert, Wilhelm Becker, Jan Huenges, Anne-Kathrin Garz, Bjoern-Oliver Gohlke, Daniel Paul Zolg, Gian Kayser, Tonu Vooder, Robert Preissner, Hannes Hahne, Neeme Tõnisson, Karl Kramer, Katharina Götze, Florian Bassermann, Judith Schlegl, Hans-Christian Ehrlich, Stephan Aiche, Axel Walch, Philipp A. Greif, Sabine Schneider, Eduard Rudolf Felder, Juergen Ruland, Guillaume Médard, Irmela Jeremias, Karsten Spiekermann, and Bernhard Kuster. The target landscape of clinical kinase drugs. *Science (New York, N.Y.)*, 358(6367): eaan4368, 2017. ISSN 1095-9203 0036-8075. doi: 10.1126/science.aan4368. URL https://pubmed.ncbi.nlm.nih.gov/29191878https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6542668/. 29191878[pmid] PMC6542668[pmcid] 358/6367/eaan4368[PII].

Ann Lin, Christopher J. Giuliano, Ann Palladino, Kristen M. John, Connor Abramowicz, Monet Lou Yuan, Erin L. Sausville, Devon A. Lukow, Luwei Liu, Alexander R. Chait, Zachary C. Galluzzo, Clara Tucker, and Jason M. Sheltzer. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science translational medicine*, 11(509):eaaw8412, 2019. ISSN 1946-6242 1946-6234. doi: 10.1126/scitranslmed.aaw8412. URL https://pubmed.ncbi.nlm.nih.gov/31511426https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7717492/. 31511426[pmid] PMC7717492[pmcid] 11/509/eaaw8412[PII].

T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan, P. L. Harris, S. M. Haserlat, J. G. Supko, F. G. Haluska, D. N. Louis, D. C. Christiani, J. Settleman, and D. A. Haber. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med*, 350(21):2129–39, 2004. ISSN 1533-4406 (Electronic) 0028-4793 (Linking). doi: 10.1056/NEJMoa040938. URL https://www.ncbi.nlm.nih.gov/pubmed/15118073.

W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, E. Mardis, D. Kupfer, R. Wilson, M. Kris, and H. Varmus. Egf receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc Natl Acad Sci U S A*, 101(36):13306–11, 2004. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.0405220101405220101[pii]. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15329413.

L. V. Sequist, J. C. Yang, N. Yamamoto, K. O'Byrne, V. Hirsh, T. Mok, S. L. Geater, S. Orlov, C. M. Tsai, M. Boyer, W. C. Su, J. Bennouna, T. Kato, V. Gorbunova, K. H. Lee, R. Shah, D. Massey, V. Zazulina, M. Shahidi, and M. Schuler. Phase iii study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with egfr mutations. *J Clin Oncol*, 31(27):3327–34, 2013. ISSN 1527-7755 (Electronic) 0732-183X (Linking). doi: 10.1200/JCO.2012.44.2806. URL https://www.ncbi.nlm.nih.gov/pubmed/23816960.

R. Rosell, E. Carcereny, R. Gervais, A. Vergnenegre, B. Massuti, E. Felip, R. Palmero, R. Garcia-Gomez, C. Pallares, J. M. Sanchez, R. Porta, M. Cobo, P. Garrido, F. Longo, T. Moran, A. Insa, F. De Marinis, R. Corre, I. Bover, A. Illiano, E. Dansin, J. de Castro, M. Milella, N. Reguart, G. Altavilla, U. Jimenez, M. Provencio, M. A. Moreno, J. Terrasa, J. Munoz-Langa, J. Valdivia, D. Isla, M. Domine, O. Molinier, J. Mazieres, N. Baize, R. Garcia-Campelo, G. Robinet, D. Rodriguez-Abreu, G. Lopez-Vivanco, V. Gebbia, L. Ferrera-Delgado, P. Bombaron, R. Bernabe, A. Bearz, A. Artal, E. Cortesi, C. Rolfo, M. Sanchez-Ronco, A. Drozdowskyj, C. Queralt, I. de Aguirre, J. L. Ramirez, J. J. Sanchez, M. A. Molina, M. Taron, L. Paz-Ares, Pneumo-Cancerologie Spanish Lung Cancer Group in collaboration with Groupe Francais de, and Toracica Associazione Italiana Oncologia. Erlotinib versus standard chemotherapy as first-line treatment for european patients with advanced egfr mutation-positive non-small-cell lung cancer (eurtac): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol*, 13(3):239–46, 2012. ISSN 1474-5488 (Electronic) 1470-2045 (Linking). doi: 10.1016/S1470-2045(11)70393-X. URL https://www.ncbi.nlm.nih.gov/pubmed/22285168.

T. Mitsudomi, S. Morita, Y. Yatabe, S. Negoro, I. Okamoto, J. Tsurutani, T. Seto, M. Satouchi, H. Tada, T. Hirashima, K. Asami, N. Katakami, M. Takada, H. Yoshioka, K. Shibata, S. Kudoh, E. Shimizu, H. Saito, S. Toyooka, K. Nakagawa, M. Fukuoka, and Group West Japan Oncology. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (wjtog3405): an open label, randomised phase 3 trial. *Lancet Oncol*, 11(2):121–8, 2010. ISSN 1474-5488 (Electronic) 1470-2045 (Linking). doi: 10.1016/S1470-2045(09)70364-X. URL https://www.ncbi.nlm.nih.gov/pubmed/20022809.

M. Maemondo, A. Inoue, K. Kobayashi, S. Sugawara, S. Oizumi, H. Isobe, A. Gemma, M. Harada, H. Yoshizawa, I. Kinoshita, Y. Fujita, S. Okinaga, H. Hirano, K. Yoshimori, T. Harada, T. Ogura, M. Ando, H. Miyazawa, T. Tanaka, Y. Saijo, K. Hagiwara, S. Morita, T. Nukiwa, and Group North-East Japan Study. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated egfr. *N Engl J Med*, 362(25):2380–8, 2010. ISSN 1533-4406 (Electronic) 0028-4793 (Linking). doi: 10.1056/NEJMoa0909530. URL https://www.ncbi.nlm.nih.gov/pubmed/20573926.

G. R. Oxnard, Y. Hu, K. F. Mileham, H. Husain, D. B. Costa, P. Tracy, N. Feeney, L. M. Sholl, S. E. Dahlberg, A. J. Redig, D. J. Kwiatkowski, M. S. Rabin, C. P. Paweletz, K. S. Thress, and P. A. Janne. Assessment of resistance mechanisms and clinical implications in patients with egfr t790m-positive lung cancer and acquired resistance to osimertinib. *JAMA Oncol*, 4(11):1527–1534, 2018. ISSN 2374-2445 (Electronic) 2374-2437 (Linking). doi: 10.1001/jamaoncol.2018.2969. URL https://www.ncbi.nlm.nih.gov/pubmed/30073261.

E. L. Stewart, S. Z. Tan, G. Liu, and M. S. Tsao. Known and putative mechanisms of resistance to egfr targeted therapies in nsclc patients with egfr mutations-a review. *Transl Lung Cancer Res*, 4(1):67–81, 2015. ISSN 2218-6751 (Print) 2218-6751 (Linking). doi: 10.3978/j.issn.2218-6751.2014.11.06. URL https://www.ncbi.nlm.nih.gov/pubmed/25806347.

H. A. Yu, M. E. Arcila, N. Rekhtman, C. S. Sima, M. F. Zakowski, W. Pao, M. G. Kris, V. A. Miller, M. Ladanyi, and G. J. Riely. Analysis of tumor specimens at the time of acquired resistance to egfr-tki therapy in 155 patients with egfr-mutant lung cancers. *Clin Cancer Res*, 19(8):2240–7, 2013. ISSN 1078-0432 (Print) 1078-0432 (Linking). doi: 10.1158/1078-0432.CCR-12-2246. URL https://www.ncbi.nlm.nih.gov/pubmed/23470965.

J. C. Yang, M. J. Ahn, D. W. Kim, S. S. Ramalingam, L. V. Sequist, W. C. Su, S. W. Kim, J. H. Kim, D. Planchard, E. Felip, F. Blackhall, D. Haggstrom, K. Yoh, S. Novello, K. Gold, T. Hirashima, C. C. Lin, H. Mann, M. Cantarini, S. Ghiorghiu, and P. A. Janne. Osimertinib in pretreated t790m-positive advanced non-small-cell lung cancer: Aura study phase ii extension component. *J Clin Oncol*, 35(12): 1288–1296, 2017. ISSN 1527-7755 (Electronic) 0732-183X (Linking). doi: 10.1200/JCO.2016.70.3223. URL https://www.ncbi.nlm.nih.gov/pubmed/28221867.

J. C. Soria, Y. Ohe, J. Vansteenkiste, T. Reungwetwattana, B. Chewaskulyong, K. H. Lee, A. Dechaphunkul, F. Imamura, N. Nogami, T. Kurata, I. Okamoto, C. Zhou, B. C. Cho, Y. Cheng, E. K. Cho, P. J. Voon, D. Planchard, W. C. Su, J. E. Gray, S. M. Lee, R. Hodge, M. Marotti, Y. Rukazenkov, S. S. Ramalingam, and Flaura Investigators. Osimertinib in untreated egfr-mutated advanced non-small-cell lung cancer. *N Engl J Med*, 378(2):113–125, 2018. ISSN 1533-4406 (Electronic) 0028-4793 (Linking). doi: 10.1056/NEJMoa1713137. URL https://www.ncbi.nlm.nih.gov/pubmed/29151359.

Han J Ahn M Ramalingam SS John T Okamoto I Yang JC Bulusu KC Laus G Collins B Barrett JC Chmielecki J Mok TS Papadimitrakopoulou VA, Wu Y. Analysis of resistance mechanisms to osimertinib in patients with egfr t790m advanced nsclc from the aura3 study. *Ann Oncol*, 29(Suppl$_8$), 2018.

Zhou C Ohe Y Imamura F Cho BC Lin M Majem M Shah R Rukazenkov Y Todd A Markovets A Barrett FC Chmielecki J Gray J Ramalingam SS, Cheng Y. Mechanisms of acquired resistance to first-line osimertinib: preliminary data from the phase iii flaura study. *Ann Oncol*, 29(Suppl$_8$), 2018.

Z. Piotrowska, H. Isozaki, J. K. Lennerz, J. F. Gainor, I. T. Lennes, V. W. Zhu, N. Marcoux, M. K. Banwait, S. R. Digumarthy, W. Su, S. Yoda, A. K. Riley, V. Nangia, J. J. Lin, R. J. Nagy, R. B. Lanman, D. Dias-Santagata, M. Mino-Kenudson, A. J. Iafrate, R. S. Heist, A. T. Shaw, E. K. Evans, C. Clifford, S. I. Ou, B. Wolf, A. N. Hata, and L. V. Sequist. Landscape of acquired resistance to osimertinib in egfr-mutant nsclc and clinical validation of combined egfr and ret inhibition with osimertinib and blu-667 for acquired ret fusion. *Cancer Discov*, 2018. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). doi: 10.1158/2159-8290.CD-18-1022. URL https://www.ncbi.nlm.nih.gov/pubmed/30257958.

S. S. Ramalingam, J. C. Yang, C. K. Lee, T. Kurata, D. W. Kim, T. John, N. Nogami, Y. Ohe, H. Mann, Y. Rukazenkov, S. Ghiorghiu, D. Stetson, A. Markovets, J. C. Barrett, K. S. Thress, and P. A. Janne. Osimertinib as first-line treatment of egfr mutation-positive advanced non-small-cell lung cancer. *J Clin Oncol*, 36 (9):841–849, 2018. ISSN 1527-7755 (Electronic) 0732-183X (Linking). doi: 10.1200/JCO.2017.74.7576. URL https://www.ncbi.nlm.nih.gov/pubmed/28841389.

K. S. Thress, C. P. Paweletz, E. Felip, B. C. Cho, D. Stetson, B. Dougherty, Z. Lai, A. Markovets, A. Vivancos, Y. Kuang, D. Ercan, S. E. Matthews, M. Cantarini, J. C. Barrett, P. A. Janne, and G. R. Oxnard. Acquired egfr c797s mutation mediates resistance to azd9291 in non-small cell lung cancer harboring egfr t790m. *Nat*

Med, 21(6):560–2, 2015. ISSN 1546-170X (Electronic) 1078-8956 (Linking). doi: 10.1038/nm.3854. URL http://www.ncbi.nlm.nih.gov/pubmed/25939061.

Y. Yosaatmadja, S. Silva, J. M. Dickson, A. V. Patterson, J. B. Smaill, J. U. Flanagan, M. J. McKeage, and C. J. Squire. Binding mode of the breakthrough inhibitor azd9291 to epidermal growth factor receptor revealed. *J Struct Biol*, 192(3):539–544, 2015. ISSN 1095-8657 (Electronic) 1047-8477 (Linking). doi: 10.1016/j.jsb.2015.10.018. URL https://www.ncbi.nlm.nih.gov/pubmed/26522274.

A. Oztan, S. Fischer, A. B. Schrock, R. L. Erlich, C. M. Lovly, P. J. Stephens, J. S. Ross, V. Miller, S. M. Ali, S. I. Ou, and L. E. Raez. Emergence of egfr g724s mutation in egfr-mutant lung adenocarcinoma post progression on osimertinib. *Lung Cancer*, 111:84–87, 2017. ISSN 1872-8332 (Electronic) 0169-5002 (Linking). doi: 10.1016/j.lungcan.2017.07.002. URL http://www.ncbi.nlm.nih.gov/pubmed/28838405.

N. Peled, L. C. Roisman, B. Miron, R. Pfeffer, R. B. Lanman, M. Ilouze, A. Dvir, L. Soussan-Gutman, F. Barlesi, G. Tarcic, O. Edelheit, D. Gandara, and Y. Elkabetz. Subclonal therapy by two egfr tkis guided by sequential plasma cell-free dna in egfr-mutated lung cancer. *J Thorac Oncol*, 12(7):e81–e84, 2017. ISSN 1556-1380 (Electronic) 1556-0864 (Linking). doi: 10.1016/j.jtho.2017.02.023. URL http://www.ncbi.nlm.nih.gov/pubmed/28286242.

J. Fassunke, F. Muller, M. Keul, S. Michels, M. A. Dammert, A. Schmitt, D. Plenker, J. Lategahn, C. Heydt, J. Bragelmann, H. L. Tumbrink, Y. Alber, S. Klein, A. Heimsoeth, I. Dahmen, R. N. Fischer, M. Scheffler, M. A. Ihle, V. Priesner, A. H. Scheel, S. Wagener, A. Kron, K. Frank, K. Garbert, T. Persigehl, M. Pusken, S. Haneder, B. Schaaf, E. Rodermann, W. Engel-Riedel, E. Felip, E. F. Smit, S. Merkelbach-Bruse, H. C. Reinhardt, S. M. Kast, J. Wolf, D. Rauh, R. Buttner, and M. L. Sos. Overcoming egfr(g724s)-mediated osimertinib resistance through unique binding characteristics of second-generation egfr inhibitors. *Nat Commun*, 9(1):4655, 2018. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-018-07078-0. URL https://www.ncbi.nlm.nih.gov/pubmed/30405134.

Z. Yu, T. J. Boggon, S. Kobayashi, C. Jin, P. C. Ma, A. Dowlati, J. A. Kern, D. G. Tenen, and B. Halmos. Resistance to an irreversible epidermal growth factor receptor (egfr) inhibitor in egfr-mutant lung cancer reveals novel treatment strategies. *Cancer Res*, 67(21):10417–27, 2007. ISSN 1538-7445 (Electronic) 0008-5472 (Linking). doi: 10.1158/0008-5472.CAN-07-1248. URL https://www.ncbi.nlm.nih.gov/pubmed/17974985.

D. Ercan, H. G. Choi, C. H. Yun, M. Capelletti, T. Xie, M. J. Eck, N. S. Gray, and P. A. Janne. Egfr mutations and resistance to irreversible pyrimidine-based egfr inhibitors. *Clin Cancer Res*, 21(17):3913–23, 2015. ISSN 1078-0432 (Print) 1078-0432 (Linking). doi: 10.1158/1078-0432.CCR-14-2789. URL https://www.ncbi.nlm.nih.gov/pubmed/25948633.

Y. Miao and J. A. McCammon. Gaussian accelerated molecular dynamics: Theory, implementation, and applications. *Annu Rep Comput Chem*, 13:231–278, 2017. ISSN 1574-1400 (Print) 1574-1400 (Linking). doi: 10.1016/bs.arcc.2017.06.005. URL https://www.ncbi.nlm.nih.gov/pubmed/29720925.

Y. Miao, V. A. Feher, and J. A. McCammon. Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *J Chem Theory Comput*, 11(8):3584–3595, 2015. ISSN 1549-9626 (Electronic) 1549-9618 (Linking). doi: 10.1021/acs.jctc.5b00436. URL https://www.ncbi.nlm.nih.gov/pubmed/26300708.

J. Farmer, F. Kanwal, N. Nikulsin, M. C. B. Tsilimigras, and D. J. Jacobs. Statistical measures to quantify similarity between molecular dynamics simulation trajectories. *Entropy*, 19(12), 2017. ISSN 1099-4300. doi: 10.3390/e19120646. URL ⟨GotoISI⟩://WOS:000419007900015.

F. Solca, G. Dahl, A. Zoephel, G. Bader, M. Sanderson, C. Klein, O. Kraemer, F. Himmelsbach, E. Haaksma, and G. R. Adolf. Target binding properties and cellular activity of afatinib (bibw 2992), an irreversible erbb family blocker. *J Pharmacol Exp Ther*, 343(2):342–50, 2012. ISSN 1521-0103 (Electronic) 0022-3565 (Linking). doi: 10.1124/jpet.112.197756. URL https://www.ncbi.nlm.nih.gov/pubmed/22888144.

Y. L. Zhang, J. Q. Yuan, K. F. Wang, X. H. Fu, X. R. Han, D. Threapleton, Z. Y. Yang, C. Mao, and J. L. Tang. The prevalence of egfr mutation in patients with non-small cell lung cancer: a systematic review and

meta-analysis. *Oncotarget*, 7(48):78985–78993, 2016. ISSN 1949-2553 (Electronic) 1949-2553 (Linking). doi: 10.18632/oncotarget.12587. URL https://www.ncbi.nlm.nih.gov/pubmed/27738317.

J. Cho, A. J. Bass, M. S. Lawrence, K. Cibulskis, A. Cho, S. N. Lee, M. Yamauchi, N. Wagle, P. Pochanard, N. Kim, A. K. Park, J. Won, H. S. Hur, H. Greulich, S. Ogino, C. Sougnez, D. Voet, J. Tabernero, J. Jimenez, J. Baselga, S. B. Gabriel, E. S. Lander, G. Getz, M. J. Eck, W. Y. Park, and M. Meyerson. Colon cancer-derived oncogenic egfr g724s mutant identified by whole genome sequence analysis is dependent on asymmetric dimerization and sensitive to cetuximab. *Mol Cancer*, 13:141, 2014. ISSN 1476-4598 (Electronic) 1476-4598 (Linking). doi: 10.1186/1476-4598-13-141. URL http://www.ncbi.nlm.nih.gov/pubmed/24894453.

Y. Kobayashi and T. Mitsudomi. Not all epidermal growth factor receptor mutations in lung cancer are created equal: Perspectives for individualized treatment strategy. *Cancer Sci*, 107(9):1179–86, 2016. ISSN 1349-7006 (Electronic) 1347-9032 (Linking). doi: 10.1111/cas.12996. URL https://www.ncbi.nlm.nih.gov/pubmed/27323238.

X. Zhang, J. Gureasko, K. Shen, P. A. Cole, and J. Kuriyan. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell*, 125(6):1137–49, 2006. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi: 10.1016/j.cell.2006.05.013. URL http://www.ncbi.nlm.nih.gov/pubmed/16777603http://ac.els-cdn.com/S0092867406005848/1-s2.0-S0092867406005848-main.pdf?_tid=2c901dde-ee7c-11e4-8560-00000aacb360&acdnat=1430317962_0436856eb995029c4b392000b157dfe4.

Y. Shan, M. P. Eastwood, X. Zhang, E. T. Kim, A. Arkhipov, R. O. Dror, J. Jumper, J. Kuriyan, and D. E. Shaw. Oncogenic mutations counteract intrinsic disorder in the egfr kinase and promote receptor dimerization. *Cell*, 149(4):860–70, 2012. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2012.02.063. URL https://www.ncbi.nlm.nih.gov/pubmed/22579287.

C. L. McClendon, A. P. Kornev, M. K. Gilson, and S. S. Taylor. Dynamic architecture of a protein kinase. *Proc Natl Acad Sci U S A*, 111(43):E4623–31, 2014. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1418402111. URL https://www.ncbi.nlm.nih.gov/pubmed/25319261.

D. A. Cross, S. E. Ashton, S. Ghiorghiu, C. Eberlein, C. A. Nebhan, P. J. Spitzler, J. P. Orme, M. R. Finlay, R. A. Ward, M. J. Mellor, G. Hughes, A. Rahi, V. N. Jacobs, M. Red Brewer, E. Ichihara, J. Sun, H. Jin, P. Ballard, K. Al-Kadhimi, R. Rowlinson, T. Klinowska, G. H. Richmond, M. Cantarini, D. W. Kim, M. R. Ranson, and W. Pao. Azd9291, an irreversible egfr tki, overcomes t790m-mediated resistance to egfr inhibitors in lung cancer. *Cancer Discov*, 4(9):1046–61, 2014. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). doi: 10.1158/2159-8290.CD-14-0337. URL https://www.ncbi.nlm.nih.gov/pubmed/24893891.

J. J. Lin, G. J. Riely, and A. T. Shaw. Targeting alk: Precision medicine takes on drug resistance. *Cancer Discov*, 7(2):137–155, 2017. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). doi: 10.1158/2159-8290.CD-16-1123. URL https://www.ncbi.nlm.nih.gov/pubmed/28122866.

A. T. Shaw and J. A. Engelman. Alk in lung cancer: past, present, and future. *J Clin Oncol*, 31(8):1105–11, 2013. ISSN 1527-7755 (Electronic) 0732-183X (Linking). doi: 10.1200/JCO.2012.44.5353. URL https://www.ncbi.nlm.nih.gov/pubmed/23401436.

J. F. Gainor, L. Dardaei, S. Yoda, L. Friboulet, I. Leshchiner, R. Katayama, I. Dagogo-Jack, S. Gadgeel, K. Schultz, M. Singh, E. Chin, M. Parks, D. Lee, R. H. DiCecca, E. Lockerman, T. Huynh, J. Logan, L. L. Ritterhouse, L. P. Le, A. Muniappan, S. Digumarthy, C. Channick, C. Keyes, G. Getz, D. Dias-Santagata, R. S. Heist, J. Lennerz, L. V. Sequist, C. H. Benes, A. J. Iafrate, M. Mino-Kenudson, J. A. Engelman, and A. T. Shaw. Molecular mechanisms of resistance to first- and second-generation alk inhibitors in alk-rearranged lung cancer. *Cancer Discov*, 6(10):1118–1133, 2016. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). doi: 10.1158/2159-8290.CD-16-0596. URL https://www.ncbi.nlm.nih.gov/pubmed/27432227.

J. J. Lin, V. W. Zhu, S. Yoda, B. Y. Yeap, A. B. Schrock, I. Dagogo-Jack, N. A. Jessop, G. Y. Jiang, L. P. Le, K. Gowen, P. J. Stephens, J. S. Ross, S. M. Ali, V. A. Miller, M. L. Johnson, C. M. Lovly, A. N. Hata, J. F. Gainor, A. J. Iafrate, A. T. Shaw, and S. I. Ou. Impact of eml4-alk variant on resistance mechanisms and clinical outcomes in alk-positive lung cancer. *J Clin Oncol*, 36(12):1199–1206, 2018. ISSN 1527-7755

(Electronic) 0732-183X (Linking). doi: 10.1200/JCO.2017.76.2294. URL https://www.ncbi.nlm.nih.gov/pubmed/29373100.

S. S. Ramalingam, J. Vansteenkiste, D. Planchard, B. C. Cho, J. E. Gray, Y. Ohe, C. Zhou, T. Reungwetwattana, Y. Cheng, B. Chewaskulyong, R. Shah, M. Cobo, K. H. Lee, P. Cheema, M. Tiseo, T. John, M. C. Lin, F. Imamura, T. Kurata, A. Todd, R. Hodge, M. Saggese, Y. Rukazenkov, J. C. Soria, and Flaura Investigators. Overall survival with osimertinib in untreated, egfr-mutated advanced nsclc. *N Engl J Med*, 382(1):41–50, 2020. ISSN 1533-4406 (Electronic) 0028-4793 (Linking). doi: 10.1056/NEJMoa1913662. URL https://www.ncbi.nlm.nih.gov/pubmed/31751012.

Mai He, Marzia Capelletti, Khedoudja Nafa, Cai-Hong Yun, Maria E. Arcila, Vincent A. Miller, Michelle S. Ginsberg, Binsheng Zhao, Mark G. Kris, Michael J. Eck, Pasi A. Jänne, Marc Ladanyi, and Geoffrey R. Oxnard. lt;emgt;egfrlt;/emgt; exon 19 insertions: A new family of sensitizing lt;emgt;egfrlt;/emgt; mutations in lung adenocarcinoma. *Clinical Cancer Research*, 18(6):1790, 2012. doi: 10.1158/1078-0432.ccr-11-2361. URL http://clincancerres.aacrjournals.org/content/18/6/1790.abstract.

Takayuki Kosaka, Junko Tanizaki, Raymond M. Paranal, Hideki Endoh, Christine Lydon, Marzia Capelletti, Claire E. Repellin, Jihyun Choi, Atsuko Ogino, Antonio Calles, Dalia Ercan, Amanda J. Redig, Magda Bahcall, Geoffrey R. Oxnard, Michael J. Eck, and Pasi A. Jänne. Response heterogeneity of egfr and her2 exon 20 insertions to covalent egfr and her2 inhibitors. *Cancer Research*, 77(10):2712, 2017. doi: 10.1158/0008-5472.can-16-3404. URL http://cancerres.aacrjournals.org/content/77/10/2712.abstract.

J. Naidoo, C. S. Sima, K. Rodriguez, N. Busby, K. Nafa, M. Ladanyi, G. J. Riely, M. G. Kris, M. E. Arcila, and H. A. Yu. Epidermal growth factor receptor exon 20 insertions in advanced lung adenocarcinomas: Clinical outcomes and response to erlotinib. *Cancer*, 121(18):3212–3220, 2015. ISSN 1097-0142 0008-543X. doi: 10.1002/cncr.29493. URL https://pubmed.ncbi.nlm.nih.gov/26096453https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4807634/.

H. Yasuda, S. Kobayashi, and D. B. Costa. Egfr exon 20 insertion mutations in non-small-cell lung cancer: preclinical data and clinical implications. *Lancet Oncol*, 13(1):e23–31, 2012. ISSN 1474-5488 (Electronic) 1470-2045 (Linking). doi: 10.1016/S1470-2045(11)70129-2. URL https://www.ncbi.nlm.nih.gov/pubmed/21764376.

Hiroyuki Yasuda, Eunyoung Park, Cai-Hong Yun, Natasha J. Sng, Antonio R. Lucena-Araujo, Wee-Lee Yeo, Mark S. Huberman, David W. Cohen, Sohei Nakayama, Kota Ishioka, Norihiro Yamaguchi, Megan Hanna, Geoffrey R. Oxnard, Christopher S. Lathan, Teresa Moran, Lecia V. Sequist, Jamie E. Chaft, Gregory J. Riely, Maria E. Arcila, Ross A. Soo, Matthew Meyerson, Michael J. Eck, Susumu S. Kobayashi, and Daniel B. Costa. Structural, biochemical, and clinical characterization of epidermal growth factor receptor (egfr) exon 20 insertion mutations in lung cancer. *Science Translational Medicine*, 5(216):216ra177, 2013. doi: 10.1126/scitranslmed.3007205. URL http://stm.sciencemag.org/content/5/216/216ra177.abstract.

Zheng Ruan and Natarajan Kannan. Altered conformational landscape and dimerization dependency underpins the activation of egfr by c-4 loop insertion mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 115(35):E8162–E8171, 2018. ISSN 1091-6490 0027-8424. doi: 10.1073/pnas.1803152115. URL https://pubmed.ncbi.nlm.nih.gov/30104348https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6126729/.

Francois Gonzalvez, Sylvie Vincent, Theresa E. Baker, Alexandra E. Gould, Shuai Li, Scott D. Wardwell, Sara Nadworny, Yaoyu Ning, Sen Zhang, Wei-Sheng Huang, Yongbo Hu, Feng Li, Matthew T. Greenfield, Stephan G. Zech, Biplab Das, Narayana I. Narasimhan, Tim Clackson, David Dalgarno, William C. Shakespeare, Michael Fitzgerald, Johara Chouitar, Robert J. Griffin, Shengwu Liu, Kwok-Kin Wong, Xiaotian Zhu, and Victor M. Rivera. Mobocertinib (tak-788): A targeted inhibitor of egfr exon 20 insertion mutants in non-small cell lung cancer. *Cancer Discovery*, page candisc.1683.2020, 2021. doi: 10.1158/2159-8290.cd-20-1683. URL http://cancerdiscovery.aacrjournals.org/content/early/2021/02/24/2159-8290.CD-20-1683.abstract.

Gregory J. Riely, Joel W. Neal, D. Ross Camidge, Alexander I. Spira, Zofia Piotrowska, Daniel B. Costa, Anne S. Tsao, Jyoti D. Patel, Shirish M. Gadgeel, Lyudmila Bazhenova, Viola W. Zhu, Howard L. West, Tarek Mekhail, Ryan D. Gentzler, Danny Nguyen, Sylvie Vincent, Steven Zhang, Jianchang Lin, Veronica Bunn, Shu Jin, Shuanglian Li, and Pasi A. Janne. Activity and safety of mobocertinib (tak-788) in previously treated non-small cell lung cancer with egfr exon 20 insertion mutations from a phase 1/2 trial. *Cancer Discovery*, page candisc.1598.2020, 2021. doi: 10.1158/2159-8290.cd-20-1598. URL http://cancerdiscovery.aacrjournals.org/content/early/2021/02/24/2159-8290.CD-20-1598.abstract.

Jaebong Jang, Jieun Son, Eunyoung Park, Takayuki Kosaka, Jamie A. Saxon, Dries J. H. DeClercq, Hwan Geun Choi, Junko Tanizaki, Michael J. Eck, Pasi A. Jänne, and Nathanael S. Gray. Discovery of a highly potent and broadly effective epidermal growth factor receptor and her2 exon 20 insertion mutant inhibitor. *Angewandte Chemie International Edition*, 57(36):11629–11633, 2018. ISSN 1433-7851. doi: https://doi.org/10.1002/anie.201805187. URL https://doi.org/10.1002/anie.201805187.

Nahomi Tokudome, Yasuhiro Koh, Hiroaki Akamatsu, Daichi Fujimoto, Isamu Okamoto, Kazuhiko Nakagawa, Toyoaki Hida, Fumio Imamura, Satoshi Morita, and Nobuyuki Yamamoto. Differential significance of molecular subtypes which were classified into egfr exon 19 deletion on the first line afatinib monotherapy. *BMC Cancer*, 20(1):103–103, 2020. ISSN 1471-2407. doi: 10.1186/s12885-020-6593-1. URL https://pubmed.ncbi.nlm.nih.gov/32028909https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7006223/.

Chao Zhao, Tao Jiang, Jiayu Li, Yan Wang, Chunxia Su, Xiaoxia Chen, Shengxiang Ren, Xuefei Li, and Caicun Zhou. The impact of egfr exon 19 deletion subtypes on clinical outcomes in non-small cell lung cancer. *Translational lung cancer research*, 9(4):1149–1158, 2020. ISSN 2218-6751 2226-4477. doi: 10.21037/tlcr-19-359. URL https://pubmed.ncbi.nlm.nih.gov/32953493https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7481579/.

Chun-Wei Xu, Lei Lei, Wen-Xian Wang, Li Lin, You-Cai Zhu, Hong Wang, Li-Yun Miao, Li-Ping Wang, Wu Zhuang, Mei-Yu Fang, Tang-Feng Lv, and Yong Song. Molecular characteristics and clinical outcomes of egfr exon 19 c-helix deletion in non-small cell lung cancer and response to egfr tkis. *Translational Oncology*, 13(9):100791–100791, 2020. ISSN 1936-5233. doi: 10.1016/j.tranon.2020.100791. URL https://pubmed.ncbi.nlm.nih.gov/32492620https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7264750/.

Kuei-Pin Chung, Shang-Gin Wu, Jenn-Yu Wu, James Chih-Hsin Yang, Chong-Jen Yu, Pin-Fei Wei, Jin-Yuan Shih, and Pan-Chyr Yang. Clinical outcomes in non–small cell lung cancers harboring different exon 19 deletions in egfr. *Clinical Cancer Research*, 18(12):3470, 2012. doi: 10.1158/1078-0432.ccr-11-2353. URL http://clincancerres.aacrjournals.org/content/18/12/3470.abstract.

Jian Su, Wenzhao Zhong, Xuchao Zhang, Ying Huang, Honghong Yan, Jinji Yang, Zhongyi Dong, Zhi Xie, Qing Zhou, Xiaosui Huang, Danxia Lu, Wenqing Yan, and Yi-Long Wu. Molecular characteristics and clinical outcomes of egfr exon 19 indel subtypes to egfr tkis in nsclc patients. *Oncotarget*, 8(67):111246–111257, 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.22768. URL https://pubmed.ncbi.nlm.nih.gov/29340050https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5762318/.

Tyler Stewart, Anna Truini, Michelle DeVeaux, Daniel Zelterman, Zenta Walther, Anna Wurtz, Scott N. Gettinger, Katerina A. Politi, and Sarah B. Goldberg. Differential outcomes in patients with uncommon egfr exon 19 mutations. *Journal of Clinical Oncology*, $36(15_{suppl}) : 9056 - -9056, 2018. ISSN 0732 - 183X. doi : .$ URL https://doi.org/10.1200/JCO.2018.36.15_suppl.9056.

J. Cho, L. Chen, N. Sangji, T. Okabe, K. Yonesaka, J. M. Francis, R. J. Flavin, W. Johnson, J. Kwon, S. Yu, H. Greulich, B. E. Johnson, M. J. Eck, P. A. Janne, K. K. Wong, and M. Meyerson. Cetuximab response of lung cancer-derived egf receptor mutants is associated with asymmetric dimerization. *Cancer Res*, 73 (22):6770–9, 2013. ISSN 1538-7445 (Electronic) 0008-5472 (Linking). 10.1158/0008-5472.CAN-13-1145. URL https://www.ncbi.nlm.nih.gov/pubmed/24063894.

Heidi Greulich, Tzu-Hsiu Chen, Whei Feng, Pasi A. Jänne, James V. Alvarez, Mauro Zappaterra, Sara E. Bulmer, David A. Frank, William C. Hahn, William R. Sellers, and Matthew Meyerson. Oncogenic transformation by inhibitor-sensitive and -resistant egfr mutants. *PLoS Medicine*, 2(11):e313, 2005. 10.1371/journal.pmed.0020313. URL https://doi.org/10.1371/journal.pmed.0020313.

Christopher C. Valley, Donna J. Arndt-Jovin, Narain Karedla, Mara P. Steinkamp, Alexey I. Chizhik, William S. Hlavacek, Bridget S. Wilson, Keith A. Lidke, and Diane S. Lidke. Enhanced dimerization drives ligand-independent activity of mutant epidermal growth factor receptor in lung cancer. *Molecular Biology of the Cell*, 26(22):4087–4099, 2015. ISSN 1939-4586 1059-1524. 10.1091/mbc.E15-05-0269. URL https://pubmed.ncbi.nlm.nih.gov/26337388https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4710239/.

Takafumi Okabe, Isamu Okamoto, Kenji Tamura, Masaaki Terashima, Takeshi Yoshida, Taroh Satoh, Minoru Takada, Masahiro Fukuoka, and Kazuhiko Nakagawa. Differential constitutive activation of the epidermal growth factor receptor in non–small cell lung cancer cells bearing lt;emgt;egfrlt;/emgt; gene mutation and amplification. *Cancer Research*, 67(5):2046, 2007. 10.1158/0008-5472.can-06-3339. URL http://cancerres.aacrjournals.org/content/67/5/2046.abstract.

Raffaella Sordella, Daphne W. Bell, Daniel A. Haber, and Jeffrey Settleman. Gefitinib-sensitizing lt;emgt;egfrlt;/emgt; mutations in lung cancer activate anti-apoptotic pathways. *Science*, 305(5687):1163, 2004. 10.1126/science.1101637. URL http://science.sciencemag.org/content/305/5687/1163.abstract.

Kendall D. Carey, Andrew J. Garton, Maria S. Romero, Jennifer Kahler, Stuart Thomson, Sarajane Ross, Frances Park, John D. Haley, Neil Gibson, and Mark X. Sliwkowski. Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer Research*, 66(16):8163, 2006. 10.1158/0008-5472.can-06-0453. URL http://cancerres.aacrjournals.org/content/66/16/8163.abstract.

Roseann Mulloy, Audrey Ferrand, Youngjoo Kim, Raffaella Sordella, Daphne W. Bell, Daniel A. Haber, Karen S. Anderson, and Jeffrey Settleman. Epidermal growth factor receptor mutants from human lung cancers exhibit enhanced catalytic activity and increased sensitivity to gefitinib. *Cancer Research*, 67(5): 2325, 2007. 10.1158/0008-5472.can-06-4293. URL http://cancerres.aacrjournals.org/content/67/5/2325.abstract.

Benjamin P. Brown, Yun-Kai Zhang, David Westover, Yingjun Yan, Huan Qiao, Vincent Huang, Zhenfang Du, Jarrod A. Smith, Jeffrey S. Ross, Vincent A. Miller, Siraj Ali, Lyudmila Bazhenova, Alexa B. Schrock, Jens Meiler, and Christine M. Lovly. On-target resistance to the mutant-selective egfr inhibitor osimertinib can develop in an allele-specific manner dependent on the original egfr-activating mutation. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 25(11):3341–3351, 2019a. ISSN 1557-3265 1078-0432. 10.1158/1078-0432.CCR-18-3829. URL https://pubmed.ncbi.nlm.nih.gov/30796031https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6548651/.

Aacr Project GENIE Consortium. Aacr project genie: Powering precision medicine through an international consortium. *Cancer Discov*, 7(8):818–831, 2017. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). 10.1158/2159-8290.CD-17-0151. URL https://www.ncbi.nlm.nih.gov/pubmed/28572459.

G. Cohen S Fau Carpenter, Jr. Carpenter G Fau King, L., and Jr. King, L. Epidermal growth factor-receptor-protein kinase interactions. co-purification of receptor and epidermal growth factor-enhanced phosphorylation activity. (0021-9258 (Print)), 1980.

S. R. Needham, S. K. Roberts, A. Arkhipov, V. P. Mysore, C. J. Tynan, L. C. Zanetti-Domingues, E. T. Kim, V. Losasso, D. Korovesis, M. Hirsch, D. J. Rolfe, D. T. Clarke, M. D. Winn, A. Lajevardipour, A. H. Clayton, L. J. Pike, M. Perani, P. J. Parker, Y. Shan, D. E. Shaw, and M. L. Martin-Fernandez. Egfr oligomerization organizes kinase-active dimers into competent signalling platforms. *Nat Commun*, 7:13307, 2016a. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). 10.1038/ncomms13307. URL http://www.ncbi.nlm.nih.gov/pubmed/27796308.

Y. Huang, S. Bharill, D. Karandur, S. M. Peterson, M. Marita, X. Shi, M. J. Kaliszewski, A. W. Smith, E. Y. Isacoff, and J. Kuriyan. Molecular basis for multimerization in the activation of the epidermal growth factor receptor. *Elife*, 5, 2016. ISSN 2050-084X (Electronic) 2050-084X (Linking). 10.7554/eLife.14107. URL https://www.ncbi.nlm.nih.gov/pubmed/27017828.

M. Red Brewer, C. H. Yun, D. Lai, M. A. Lemmon, M. J. Eck, and W. Pao. Mechanism for activation of mutated epidermal growth factor receptors in lung cancer. *Proc Natl Acad Sci U S A*, 110(38):E3595–604, 2013. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). 10.1073/pnas.1220050110. URL http://www.ncbi.nlm.nih.gov/pubmed/24019492.

Yibing Shan, Anton Arkhipov, Eric T. Kim, Albert C. Pan, and David E. Shaw. Transitions to catalytically inactive conformations in egfr kinase. *Proceedings of the National Academy of Sciences*, 110(18):7270, 2013. 10.1073/pnas.1220843110. URL http://www.pnas.org/content/110/18/7270.abstract.

Sheng-Kai Liang, Jen-Chung Ko, James Chih-Hsin Yang, and Jin-Yuan Shih. Afatinib is effective in the treatment of lung adenocarcinoma with uncommon ¡em¿egfr¡/em¿ p.l747p and p.l747s mutations. *Lung Cancer*, 133:103–109, 2019. ISSN 0169-5002. 10.1016/j.lungcan.2019.05.019. URL https://doi.org/10.1016/j.lungcan.2019.05.019.

Chung-Jung Tsai and Ruth Nussinov. Emerging allosteric mechanism of egfr activation in physiological and pathological contexts. *Biophysical Journal*, 117(1):5–13, 2019. ISSN 1542-0086 0006-3495. 10.1016/j.bpj.2019.05.021. URL https://pubmed.ncbi.nlm.nih.gov/31202480https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6626828/.

Zhenfang Du, Benjamin P. Brown, Soyeon Kim, Donna Ferguson, Dean C. Pavlick, Gowtham Jayakumaran, Ryma Benayed, Jean-Nicolas Gallant, Yun-Kai Zhang, Yingjun Yan, Monica Red-Brewer, Siraj M. Ali, Alexa B. Schrock, Ahmet Zehir, Marc Ladanyi, Adam W. Smith, Jens Meiler, and Christine M. Lovly. Structure–function analysis of oncogenic egfr kinase domain duplication reveals insights into activation and a potential approach for therapeutic targeting. *Nature Communications*, 12(1):1382, 2021. ISSN 2041-1723. 10.1038/s41467-021-21613-6. URL https://doi.org/10.1038/s41467-021-21613-6.

Yongjian Huang, Jana Ognjenović, Deepti Karandur, Alan Merk, Sriram Subramaniam, and John Kuriyan. A structural mechanism for the generation of biased agonism in the epidermal growth factor receptor. *bioRxiv*, page 2020.12.08.417006, 2020. 10.1101/2020.12.08.417006. URL http://biorxiv.org/content/early/2020/12/09/2020.12.08.417006.abstract.

A. H. Clayton, F. Walker, S. G. Orchard, C. Henderson, D. Fuchs, J. Rothacker, E. C. Nice, and A. W. Burgess. Ligand-induced dimer-tetramer transition during the activation of the cell surface epidermal growth factor receptor-a multidimensional microscopy analysis. *J Biol Chem*, 280(34):30392–9, 2005. ISSN 0021-9258 (Print) 0021-9258 (Linking). 10.1074/jbc.M504770200. URL https://www.ncbi.nlm.nih.gov/pubmed/15994331.

Anna Truini, Jacqueline H. Starrett, Tyler Stewart, Kumar Ashtekar, Zenta Walther, Anna Wurtz, David Lu, Jin H. Park, Michelle DeVeaux, Xiaoling Song, Scott Gettinger, Daniel Zelterman, Mark A. Lemmon, Sarah B. Goldberg, and Katerina Politi. The egfr exon 19 mutant l747-a750¿p exhibits distinct sensitivity to tyrosine kinase inhibitors in lung adenocarcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 25(21):6382–6391, 2019. ISSN 1557-3265 1078-0432. 10.1158/1078-0432.CCR-19-0780. URL https://pubmed.ncbi.nlm.nih.gov/31182434https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6825535/.

C. H. Yun, K. E. Mengwasser, A. V. Toms, M. S. Woo, H. Greulich, K. K. Wong, M. Meyerson, and M. J. Eck. The t790m mutation in egfr kinase causes drug resistance by increasing the affinity for atp. *Proc Natl Acad Sci U S A*, 105(6):2070–5, 2008a. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). 10.1073/pnas.0709662105. URL https://www.ncbi.nlm.nih.gov/pubmed/18227510.

Kathryn Brown, Craig Comisar, Han Witjes, John Maringwa, Rik de Greef, Karthick Vishwanathan, Mireille Cantarini, and Eugène Cox. Population pharmacokinetics and exposure-response of osimertinib in patients with non-small cell lung cancer. *British Journal of Clinical Pharmacology*, 83(6):1216–1226, 2017. ISSN 1365-2125 0306-5251. 10.1111/bcp.13223. URL https://pubmed.ncbi.nlm.nih.gov/28009438https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5427226/.

Ariella B. Hanker, Benjamin P. Brown, Jens Meiler, Arnaldo Marín, Harikrishna S. Jayanthan, Dan Ye, Chang-Ching Lin, Hiroaki Akamatsu, Kyung-Min Lee, Sumanta Chatterjee, Dhivya R. Sudhan, Alberto Servetto, Monica Red Brewer, James P. Koch, Jonathan H. Sheehan, Jie He, Alshad S. Lalani, and Carlos L. Arteaga. Co-occurring gain-of-function mutations in her2 and her3 modulate her2/her3 activation, oncogenesis, and her2 inhibitor sensitivity. *Cancer Cell*, 2021. ISSN 1535-6108. https://doi.org/10.1016/j.ccell.2021.06.001. URL https://www.sciencedirect.com/science/article/pii/S1535610821002841.

Cai-Hong Yun, Kristen E. Mengwasser, Angela V. Toms, Michele S. Woo, Heidi Greulich, Kwok-Kin Wong, Matthew Meyerson, and Michael J. Eck. The t790m mutation in egfr kinase causes drug resistance by

increasing the affinity for atp. *Proceedings of the National Academy of Sciences*, 105(6):2070, 2008b. 10.1073/pnas.0709662105. URL http://www.pnas.org/content/105/6/2070.abstract.

A. Yver. Osimertinib (azd9291)x2014;a science-driven, collaborative approach to rapid drug design and development. *Annals of Oncology*, 27(6):1165–1170, 2016. ISSN 0923-7534. 10.1093/annonc/mdw129. URL https://doi.org/10.1093/annonc/mdw129.

Qiufan Zheng, Yan Huang, Hongyun Zhao, Yunpeng Yang, Shaodong Hong, Xue Hou, Yuanyuan Zhao, Yuxiang Ma, Ting Zhou, Yaxiong Zhang, Wenfeng Fang, and Li Zhang. Egfr mutation genotypes affect efficacy and resistance mechanisms of osimertinib in t790m-positive nsclc patients. *Translational lung cancer research*, 9(3):471–483, 2020. ISSN 2218-6751 2226-4477. 10.21037/tlcr.2020.03.35. URL https://pubmed.ncbi.nlm.nih.gov/32676311https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7354104/.

S. Yoshikawa, M. Kukimoto-Niino, L. Parker, N. Handa, T. Terada, T. Fujimoto, Y. Terazawa, M. Wakiyama, M. Sato, S. Sano, T. Kobayashi, T. Tanaka, L. Chen, Z. J. Liu, B. C. Wang, M. Shirouzu, S. Kawa, K. Semba, T. Yamamoto, and S. Yokoyama. Structural basis for the altered drug sensitivities of non-small cell lung cancer-associated mutants of human epidermal growth factor receptor. *Oncogene*, 32(1):27–38, 2013. ISSN 1476-5594. 10.1038/onc.2012.21. URL https://doi.org/10.1038/onc.2012.21.

Phillip A. Schwartz, Petr Kuzmic, James Solowiej, Simon Bergqvist, Ben Bolanos, Chau Almaden, Asako Nagata, Kevin Ryan, Junli Feng, Deepak Dalvie, John C. Kath, Meirong Xu, Revati Wani, and Brion William Murray. Covalent egfr inhibitor analysis reveals importance of reversible interactions to potency and mechanisms of drug resistance. *Proceedings of the National Academy of Sciences*, 111(1):173–178, 2014. 10.1073/pnas.1313733111. URL https://www.pnas.org/content/pnas/111/1/173.full.pdf.

M. J. Kaliszewski, X. Shi, Y. Hou, R. Lingerak, S. Kim, P. Mallory, and A. W. Smith. Quantifying membrane protein oligomerization with fluorescence cross-correlation spectroscopy. *Methods*, 140-141:40–51, 2018. ISSN 1095-9130 (Electronic) 1046-2023 (Linking). 10.1016/j.ymeth.2018.02.002. URL https://www.ncbi.nlm.nih.gov/pubmed/29448037.

W. D. Comar, S. M. Schubert, B. Jastrzebska, K. Palczewski, and A. W. Smith. Time-resolved fluorescence spectroscopy measures clustering and mobility of a g protein-coupled receptor opsin in live cell membranes. *J Am Chem Soc*, 136(23):8342–9, 2014. ISSN 1520-5126 (Electronic) 0002-7863 (Linking). 10.1021/ja501948w. URL https://www.ncbi.nlm.nih.gov/pubmed/24831851.

Y. Song, F. DiMaio, R. Y. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, and D. Baker. High-resolution comparative modeling with rosettacm. *Structure*, 21(10):1735–42, 2013. ISSN 1878-4186 (Electronic) 0969-2126 (Linking). 10.1016/j.str.2013.08.005. URL http://www.ncbi.nlm.nih.gov/pubmed/24035711.

D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2018. *University of California, San Francisco.*, 2018.

3rd Miller, B. R., Jr. McGee, T. D., J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. Mmpbsa.py: An efficient program for end-state free energy calculations. *J Chem Theory Comput*, 8(9):3314–21, 2012. ISSN 1549-9618 (Print) 1549-9618 (Linking). 10.1021/ct300418h. URL https://www.ncbi.nlm.nih.gov/pubmed/26605738.

D. R. Roe and 3rd Cheatham, T. E. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput*, 9(7):3084–95, 2013. ISSN 1549-9618 (Print) 1549-9618 (Linking). 10.1021/ct400341p. URL https://www.ncbi.nlm.nih.gov/pubmed/26583988.

Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. Pyemma 2: A software package for estimation, validation, and analysis of markov models. *Journal of Chemical Theory and Computation*, 11(11):5525–5542, 2015. ISSN 1549-9618. 10.1021/acs.jctc.5b00743. URL https://doi.org/10.1021/acs.jctc.5b00743.

Satoshi Sogabe, Youichi Kawakita, Shigeru Igaki, Hidehisa Iwata, Hiroshi Miki, Douglas R. Cary, Terufumi Takagi, Shinji Takagi, Yoshikazu Ohta, and Tomoyasu Ishikawa. Structure-based approach for the discovery of pyrrolo[3,2-d]pyrimidine-based egfr t790m/l858r mutant inhibitors. *ACS medicinal chemistry letters*, 4(2):201–205, 2012. ISSN 1948-5875. 10.1021/ml300327z. URL https://pubmed.ncbi.nlm.nih.gov/24900643https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4027575/.

L. Meagher Kristin, T. Redman Luke, and A. Carlson Heather. Development of polyphosphate parameters for use with the amber force field. *Journal of Computational Chemistry*, 24(9):1016–1025, 2003. ISSN 0192-8651. 10.1002/jcc.10262. URL https://doi.org/10.1002/jcc.10262.

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017. ISSN 1549-9618. 10.1021/acs.jctc.7b00125. URL https://doi.org/10.1021/acs.jctc.7b00125https://pubs.acs.org/doi/full/10.1021/acs.jctc.7b00125.

I. S. Joung and 3rd Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B*, 112(30):9020–41, 2008. ISSN 1520-6106 (Print) 1520-5207 (Linking). 10.1021/jp8001614. URL https://www.ncbi.nlm.nih.gov/pubmed/18593145.

Alan. Grossfield. Wham: The weighted histogram analysis method.

J. J. Wheler, F. Janku, A. Naing, Y. Li, B. Stephen, R. Zinner, V. Subbiah, S. Fu, D. Karp, G. S. Falchook, A. M. Tsimberidou, S. Piha-Paul, R. Anderson, D. Ke, V. Miller, R. Yelensky, J. J. Lee, D. S. Hong, and R. Kurzrock. Cancer therapy directed by comprehensive genomic profiling: A single center study. *Cancer Res*, 76(13):3690–701, 2016. ISSN 1538-7445 (Electronic) 0008-5472 (Linking). 10.1158/0008-5472.CAN-15-3043. URL https://www.ncbi.nlm.nih.gov/pubmed/27197177.

J. S. Ross, K. Wang, L. Gay, G. A. Otto, E. White, K. Iwanik, G. Palmer, R. Yelensky, D. M. Lipson, J. Chmielecki, R. L. Erlich, A. N. Rankin, S. M. Ali, J. A. Elvin, D. Morosini, V. A. Miller, and P. J. Stephens. Comprehensive genomic profiling of carcinoma of unknown primary site: New routes to targeted therapies. *JAMA Oncol*, 1(1):40–49, 2015. ISSN 2374-2445 (Electronic) 2374-2437 (Linking). 10.1001/jamaoncol.2014.216. URL https://www.ncbi.nlm.nih.gov/pubmed/26182302.

A. Qin, A. Johnson, J. S. Ross, V. A. Miller, S. M. Ali, A. B. Schrock, and S. M. Gadgeel. Detection of known and novel fgfr fusions in non-small cell lung cancer by comprehensive genomic profiling. *J Thorac Oncol*, 14(1):54–62, 2019. ISSN 1556-1380 (Electronic) 1556-0864 (Linking). 10.1016/j.jtho.2018.09.014. URL https://www.ncbi.nlm.nih.gov/pubmed/30267839.

U. Disel, R. Madison, K. Abhishek, J. H. Chung, S. E. Trabucco, A. O. Matos, G. M. Frampton, L. A. Albacker, V. Reddy, N. Karadurmus, A. Benson, J. Webster, S. Paydas, R. Cabanillas, C. Nangia, M. A. Ozturk, S. Z. Millis, S. K. Pal, B. Wilky, E. S. Sokol, L. M. Gay, S. Soman, S. Ganesan, K. Janeway, P. J. Stephens, V. W. Zhu, S. I. Ou, C. M. Lovly, M. Gounder, A. B. Schrock, J. S. Ross, V. A. Miller, S. J. Klempner, and S. M. Ali. The pan-cancer landscape of coamplification of the tyrosine kinases kit, kdr, and pdgfra. *Oncologist*, 25(1):e39–e47, 2020. ISSN 1549-490X (Electronic) 1083-7159 (Linking). 10.1634/theoncologist.2018-0528. URL https://www.ncbi.nlm.nih.gov/pubmed/31604903.

S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, and Acmg Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet Med*, 17(5):405–24, 2015. ISSN 1530-0366 (Electronic) 1098-3600 (Linking). 10.1038/gim.2015.30. URL https://www.ncbi.nlm.nih.gov/pubmed/25741868.

M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, S. Roy, A. M. Tsimberidou, C. L. Vnencak-Jones, D. J. Wolff, A. Younes, and M. N. Nikiforova. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the association for molecular pathology, american society of clinical oncology, and college of american pathologists. *J Mol Diagn*, 19

(1):4–23, 2017. ISSN 1943-7811 (Electronic) 1525-1578 (Linking). 10.1016/j.jmoldx.2016.10.002. URL https://www.ncbi.nlm.nih.gov/pubmed/27993330.

J. N. Gallant, J. H. Sheehan, T. M. Shaver, M. Bailey, D. Lipson, R. Chandramohan, M. Red Brewer, S. J. York, M. G. Kris, J. A. Pietenpol, M. Ladanyi, V. A. Miller, S. M. Ali, J. Meiler, and C. M. Lovly. Egfr kinase domain duplication (egfr-kdd) is a novel oncogenic driver in lung cancer that is clinically responsive to afatinib. *Cancer Discov*, 5(11):1155–63, 2015. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). 10.1158/2159-8290.CD-15-0654. URL http://www.ncbi.nlm.nih.gov/pubmed/26286086.

M. Dankner, A. A. N. Rose, S. Rajkumar, P. M. Siegel, and I. R. Watson. Classifying braf alterations in cancer: new rational therapeutic strategies for actionable mutations. *Oncogene*, 37(24):3183–3199, 2018. ISSN 1476-5594 (Electronic) 0950-9232 (Linking). 10.1038/s41388-018-0171-x. URL https://www.ncbi.nlm.nih.gov/pubmed/29540830.

Z. Yao, N. M. Torres, A. Tao, Y. Gao, L. Luo, Q. Li, E. de Stanchina, O. Abdel-Wahab, D. B. Solit, P. I. Poulikakos, and N. Rosen. Braf mutants evade erk-dependent feedback by different mechanisms that determine their sensitivity to pharmacologic inhibition. *Cancer Cell*, 28(3):370–83, 2015. ISSN 1878-3686 (Electronic) 1535-6108 (Linking). 10.1016/j.ccell.2015.08.001. URL https://www.ncbi.nlm.nih.gov/pubmed/26343582.

W. Pao and J. Chmielecki. Rational, biologically based treatment of egfr-mutant non-small-cell lung cancer. *Nat Rev Cancer*, 10(11):760–74, 2010. ISSN 1474-1768 (Electronic) 1474-175X (Linking). 10.1038/nrc2947. URL https://www.ncbi.nlm.nih.gov/pubmed/20966921.

I. Vivanco, H. I. Robins, D. Rohle, C. Campos, C. Grommes, P. L. Nghiemphu, S. Kubek, B. Oldrini, M. G. Chheda, N. Yannuzzi, H. Tao, S. Zhu, A. Iwanami, D. Kuga, J. Dang, A. Pedraza, C. W. Brennan, A. Heguy, L. M. Liau, F. Lieberman, W. K. Yung, M. R. Gilbert, D. A. Reardon, J. Drappatz, P. Y. Wen, K. R. Lamborn, S. M. Chang, M. D. Prados, H. A. Fine, S. Horvath, N. Wu, A. B. Lassman, L. M. DeAngelis, W. H. Yong, J. G. Kuhn, P. S. Mischel, M. P. Mehta, T. F. Cloughesy, and I. K. Mellinghoff. Differential sensitivity of glioma- versus lung cancer-specific egfr mutations to egfr kinase inhibitors. *Cancer Discov*, 2(5):458–71, 2012. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). 10.1158/2159-8290.CD-11-0284. URL http://www.ncbi.nlm.nih.gov/pubmed/22588883.

C. W. Brennan, R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, S. H. Berman, R. Beroukhim, B. Bernard, C. J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S. A. Shukla, G. Ciriello, W. K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, D. D. Bigner, E. G. Van Meir, M. Prados, A. Sloan, K. L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, D. W. Andrews, A. Guha, M. Iacocca, B. P. O'Neill, G. Foltz, J. Myers, D. J. Weisenberger, R. Penny, R. Kucherlapati, C. M. Perou, D. N. Hayes, R. Gibbs, M. Marra, G. B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, P. W. Laird, D. Haussler, G. Getz, L. Chin, and Tcga Research Network. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–77, 2013. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). 10.1016/j.cell.2013.09.034. URL https://www.ncbi.nlm.nih.gov/pubmed/24120142.

M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M. S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. de Waal, T. Sharifnia, A. Brooks, H. Greulich, S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansen, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B. E. Johnson, P. A. Janne, V. A. Miller, W. Pao, W. D. Travis, H. I. Pass, S. B. Gabriel, E. S. Lander, R. K. Thomas, L. A. Garraway, G. Getz, and M. Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–20, 2012. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). 10.1016/j.cell.2012.08.029. URL https://www.ncbi.nlm.nih.gov/pubmed/22980975.

V. Frattini, V. Trifonov, J. M. Chan, A. Castano, M. Lia, F. Abate, S. T. Keir, A. X. Ji, P. Zoppoli, F. Niola, C. Danussi, I. Dolgalev, P. Porrati, S. Pellegatta, A. Heguy, G. Gupta, D. J. Pisapia, P. Canoll, J. N. Bruce, R. E. McLendon, H. Yan, K. Aldape, G. Finocchiaro, T. Mikkelsen, G. G. Prive, D. D. Bigner, A. Lasorella, R. Rabadan, and A. Iavarone. The integrated landscape of driver genomic alterations in glioblastoma. *Nat Genet*, 45(10):1141–9, 2013. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). 10.1038/ng.2734. URL https://www.ncbi.nlm.nih.gov/pubmed/23917401.

Network Cancer Genome Atlas Research. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–50, 2014. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). 10.1038/nature13385. URL https://www.ncbi.nlm.nih.gov/pubmed/25079552.

R. Mishra, A. B. Hanker, and J. T. Garrett. Genomic alterations of erbb receptors in cancer: clinical implications. *Oncotarget*, 8(69):114371–114392, 2017. ISSN 1949-2553 (Electronic) 1949-2553 (Linking). 10.18632/oncotarget.22825. URL https://www.ncbi.nlm.nih.gov/pubmed/29371993.

A. Zehir, R. Benayed, R. H. Shah, A. Syed, S. Middha, H. R. Kim, P. Srinivasan, J. Gao, D. Chakravarty, S. M. Devlin, M. D. Hellmann, D. A. Barron, A. M. Schram, M. Hameed, S. Dogan, D. S. Ross, J. F. Hechtman, D. F. DeLair, J. Yao, D. L. Mandelker, D. T. Cheng, R. Chandramohan, A. S. Mohanty, R. N. Ptashkin, G. Jayakumaran, M. Prasad, M. H. Syed, A. B. Rema, Z. Y. Liu, K. Nafa, L. Borsu, J. Sadowska, J. Casanova, R. Bacares, I. J. Kiecka, A. Razumova, J. B. Son, L. Stewart, T. Baldi, K. A. Mullaney, H. Al-Ahmadie, E. Vakiani, A. A. Abeshouse, A. V. Penson, P. Jonsson, N. Camacho, M. T. Chang, H. H. Won, B. E. Gross, R. Kundra, Z. J. Heins, H. W. Chen, S. Phillips, H. Zhang, J. Wang, A. Ochoa, J. Wills, M. Eubank, S. B. Thomas, S. M. Gardos, D. N. Reales, J. Galle, R. Durany, R. Cambria, W. Abida, A. Cercek, D. R. Feldman, M. M. Gounder, A. A. Hakimi, J. J. Harding, G. Iyer, Y. Y. Janjigian, E. J. Jordan, C. M. Kelly, M. A. Lowery, L. G. T. Morris, A. M. Omuro, N. Raj, P. Razavi, A. N. Shoushtari, N. Shukla, T. E. Soumerai, A. M. Varghese, R. Yaeger, J. Coleman, B. Bochner, G. J. Riely, L. B. Saltz, H. I. Scher, P. J. Sabbatini, M. E. Robson, D. S. Klimstra, B. S. Taylor, J. Baselga, N. Schultz, D. M. Hyman, M. E. Arcila, D. B. Solit, M. Ladanyi, and M. F. Berger. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med*, 23(6):703–713, 2017. ISSN 1546-170X (Electronic) 1078-8956 (Linking). 10.1038/nm.4333. URL https://www.ncbi.nlm.nih.gov/pubmed/28481359.

S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjaerde, A. Langerod, M. Ringner, S. M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. Hooijer, S. J. Jang, D. R. Jones, H. Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J. Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporte, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodriguez-Gonzalez, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van't Veer, A. Tutt, S. Knappskog, B. K. Tan, J. Jonkers, A. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. Martens, A. L. Borresen-Dale, A. L. Richardson, G. Kong, G. Thomas, and M. R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534 (7605):47–54, 2016. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). 10.1038/nature17676. URL https://www.ncbi.nlm.nih.gov/pubmed/27135926.

M. A. Lemmon and J. Schlessinger. Cell signaling by receptor tyrosine kinases. *Cell*, 141(7):1117–34, 2010. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). 10.1016/j.cell.2010.06.011. URL http://www.ncbi.nlm.nih.gov/pubmed/20602996.

R. M. Pruss and H. R. Herschman. Variants of 3t3 cells lacking mitogenic response to epidermal growth factor. *Proc Natl Acad Sci U S A*, 74(9):3918–21, 1977. ISSN 0027-8424 (Print) 0027-8424 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/302945.

R. S. Dise, M. R. Frey, R. H. Whitehead, and D. B. Polk. Epidermal growth factor stimulates rac activation through src and phosphatidylinositol 3-kinase to promote colonic epithelial cell migration. *Am J Physiol Gastrointest Liver Physiol*, 294(1):G276–85, 2008. ISSN 0193-1857 (Print) 0193-1857 (Linking). 10.1152/ajpgi.00340.2007. URL https://www.ncbi.nlm.nih.gov/pubmed/17991704.

G. M. Walton, W. S. Chen, M. G. Rosenfeld, and G. N. Gill. Analysis of deletions of the carboxyl terminus of the epidermal growth factor receptor reveals self-phosphorylation at tyrosine 992 and enhanced in vivo tyrosine phosphorylation of cell substrates. *J Biol Chem*, 265(3):1750–4, 1990. ISSN 0021-9258 (Print) 0021-9258 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/1688559.

K. Helin, T. Velu, P. Martin, W. C. Vass, G. Allevato, D. R. Lowy, and L. Beguinot. The biological activity of the human epidermal growth factor receptor is positively regulated by its c-terminal tyrosines. *Oncogene*,

6(5):825–32, 1991. ISSN 0950-9232 (Print) 0950-9232 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/1646987.

M. Red Brewer, S. H. Choi, D. Alvarado, K. Moravcevic, A. Pozzi, M. A. Lemmon, and G. Carpenter. The juxtamembrane region of the egf receptor functions as an activation domain. *Mol Cell*, 34(6):641–51, 2009. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). 10.1016/j.molcel.2009.04.034. URL http://www.ncbi.nlm.nih.gov/pubmed/19560417.

Sarah R. Needham, Selene K. Roberts, Anton Arkhipov, Venkatesh P. Mysore, Christopher J. Tynan, Laura C. Zanetti-Domingues, Eric T. Kim, Valeria Losasso, Dimitrios Korovesis, Michael Hirsch, Daniel J. Rolfe, David T. Clarke, Martyn D. Winn, Alireza Lajevardipour, Andrew H. A. Clayton, Linda J. Pike, Michela Perani, Peter J. Parker, Yibing Shan, David E. Shaw, and Marisa L. Martin-Fernandez. Egfr oligomerization organizes kinase-active dimers into competent signalling platforms. *Nature Communications*, 7:13307, 2016b. 10.1038/ncomms13307 https://www.nature.com/articles/ncomms13307supplementary-information. URL https://doi.org/10.1038/ncomms13307.

N. Jura, N. F. Endres, K. Engel, S. Deindl, R. Das, M. H. Lamers, D. E. Wemmer, X. Zhang, and J. Kuriyan. Mechanism for activation of the egf receptor catalytic domain by the juxtamembrane segment. *Cell*, 137(7):1293–307, 2009. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). 10.1016/j.cell.2009.04.025. URL http://www.ncbi.nlm.nih.gov/pubmed/19563760.

J. Wang, X. Li, X. Xue, Q. Ou, X. Wu, Y. Liang, X. Wang, M. You, Y. W. Shao, Z. Zhang, and S. Zhang. Clinical outcomes of egfr kinase domain duplication to targeted therapies in nsclc. *Int J Cancer*, 144(11):2677–2682, 2019b. ISSN 1097-0215 (Electronic) 0020-7136 (Linking). 10.1002/ijc.31895. URL https://www.ncbi.nlm.nih.gov/pubmed/30255937.

S. A. Foster, D. M. Whalen, A. Ozen, M. J. Wongchenko, J. Yin, I. Yen, G. Schaefer, J. D. Mayfield, J. Chmielecki, P. J. Stephens, L. A. Albacker, Y. Yan, K. Song, G. Hatzivassiliou, C. Eigenbrot, C. Yu, A. S. Shaw, G. Manning, N. J. Skelton, S. G. Hymowitz, and S. Malek. Activation mechanism of oncogenic deletion mutations in braf, egfr, and her2. *Cancer Cell*, 29(4):477–493, 2016. ISSN 1878-3686 (Electronic) 1535-6108 (Linking). 10.1016/j.ccell.2016.02.010. URL https://www.ncbi.nlm.nih.gov/pubmed/26996308.

J. F. Doody, Y. Wang, S. N. Patel, C. Joynes, S. P. Lee, J. Gerlak, R. L. Rolser, Y. Li, P. Steiner, R. Bassi, D. J. Hicklin, and Y. R. Hadari. Inhibitory activity of cetuximab on epidermal growth factor receptor mutations in non small cell lung cancers. *Mol Cancer Ther*, 6(10):2642–51, 2007. ISSN 1535-7163 (Print) 1535-7163 (Linking). 10.1158/1535-7163.MCT-06-0506. URL https://www.ncbi.nlm.nih.gov/pubmed/17913857.

M. Perez-Torres, M. Guix, A. Gonzalez, and C. L. Arteaga. Epidermal growth factor receptor (egfr) antibody down-regulates mutant receptors and inhibits tumors expressing egfr mutations. *J Biol Chem*, 281(52):40183–92, 2006. ISSN 0021-9258 (Print) 0021-9258 (Linking). 10.1074/jbc.M607958200. URL https://www.ncbi.nlm.nih.gov/pubmed/17082181.

C. H. Yun, T. J. Boggon, Y. Li, M. S. Woo, H. Greulich, M. Meyerson, and M. J. Eck. Structures of lung cancer-derived egfr mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell*, 11(3):217–27, 2007. ISSN 1535-6108 (Print) 1535-6108 (Linking). 10.1016/j.ccr.2006.12.017. URL http://www.ncbi.nlm.nih.gov/pubmed/17349580.

T. M. Gilmer, L. Cable, K. Alligood, D. Rusnak, G. Spehar, K. T. Gallagher, E. Woldu, H. L. Carter, A. T. Truesdale, L. Shewchuk, and E. R. Wood. Impact of common epidermal growth factor receptor and her2 variants on receptor activity and inhibition by lapatinib. *Cancer Res*, 68(2):571–9, 2008. ISSN 1538-7445 (Electronic) 0008-5472 (Linking). 10.1158/0008-5472.CAN-07-2404. URL https://www.ncbi.nlm.nih.gov/pubmed/18199554.

L. M. Sholl, B. Y. Yeap, A. J. Iafrate, A. J. Holmes-Tisch, Y. P. Chou, M. T. Wu, Y. G. Goan, L. Su, E. Benedettini, J. Yu, M. Loda, P. A. Janne, D. C. Christiani, and L. R. Chirieac. Lung adenocarcinoma with egfr amplification has distinct clinicopathologic and molecular features in never-smokers. *Cancer Res*, 69(21):8341–8, 2009. ISSN 1538-7445 (Electronic) 0008-5472 (Linking). 10.1158/0008-5472.CAN-09-2477. URL https://www.ncbi.nlm.nih.gov/pubmed/19826035.

L. Regales, Y. Gong, R. Shen, E. de Stanchina, I. Vivanco, A. Goel, J. A. Koutcher, M. Spassova, O. Ouerfelli, I. K. Mellinghoff, M. F. Zakowski, K. A. Politi, and W. Pao. Dual targeting of egfr can overcome a major

drug resistance mutation in mouse models of egfr mutant lung cancer. *J Clin Invest*, 119(10):3000–10, 2009. ISSN 1558-8238 (Electronic) 0021-9738 (Linking). 10.1172/JCI38746. URL https://www.ncbi.nlm.nih.gov/pubmed/19759520.

R. H. Whitehead, P. E. VanEeden, M. D. Noble, P. Ataliotis, and P. S. Jat. Establishment of conditionally immortalized epithelial cell lines from both colon and small intestine of adult h-2kb-tsa58 transgenic mice. *Proc Natl Acad Sci U S A*, 90(2):587–91, 1993. ISSN 0027-8424 (Print) 0027-8424 (Linking). 10.1073/pnas.90.2.587. URL https://www.ncbi.nlm.nih.gov/pubmed/7678459.

T. G. Johns, R. B. Luwor, C. Murone, F. Walker, J. Weinstock, A. A. Vitali, R. M. Perera, A. A. Jungbluth, E. Stockert, L. J. Old, E. C. Nice, A. W. Burgess, and A. M. Scott. Antitumor efficacy of cytotoxic drugs and the monoclonal antibody 806 is enhanced by the egf receptor inhibitor ag1478. *Proc Natl Acad Sci U S A*, 100(26):15871–6, 2003. ISSN 0027-8424 (Print) 0027-8424 (Linking). 10.1073/pnas.2036503100. URL https://www.ncbi.nlm.nih.gov/pubmed/14676326.

T. G. Johns, E. Stockert, G. Ritter, A. A. Jungbluth, H. J. Huang, W. K. Cavenee, F. E. Smyth, C. M. Hall, N. Watson, E. C. Nice, W. J. Gullick, L. J. Old, A. W. Burgess, and A. M. Scott. Novel monoclonal antibody specific for the de2-7 epidermal growth factor receptor (egfr) that also recognizes the egfr expressed in cells containing amplification of the egfr gene. *Int J Cancer*, 98(3):398–408, 2002. ISSN 0020-7136 (Print) 0020-7136 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/11920591.

N. F. Endres, R. Das, A. W. Smith, A. Arkhipov, E. Kovacs, Y. Huang, J. G. Pelton, Y. Shan, D. E. Shaw, D. E. Wemmer, J. T. Groves, and J. Kuriyan. Conformational coupling across the plasma membrane in activation of the egf receptor. *Cell*, 152(3):543–56, 2013. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). 10.1016/j.cell.2012.12.032. URL https://www.ncbi.nlm.nih.gov/pubmed/23374349.

S. Borowicz, M. Van Scoyk, S. Avasarala, M. K. Karuppusamy Rathinam, J. Tauler, R. K. Bikkavilli, and R. A. Winn. The soft agar colony formation assay. *J Vis Exp*, (92):e51998, 2014. ISSN 1940-087X (Electronic) 1940-087X (Linking). 10.3791/51998. URL http://www.ncbi.nlm.nih.gov/pubmed/25408172.

S. Horibata, T. V. Vo, V. Subramanian, P. R. Thompson, and S. A. Coonrod. Utilization of the soft agar colony formation assay to identify inhibitors of tumorigenicity in breast cancer cells. *J Vis Exp*, (99):e52727, 2015. ISSN 1940-087X (Electronic) 1940-087X (Linking). 10.3791/52727. URL http://www.ncbi.nlm.nih.gov/pubmed/26067809.

G. M. Frampton, A. Fichtenholtz, G. A. Otto, K. Wang, S. R. Downing, J. He, M. Schnall-Levin, J. White, E. M. Sanford, P. An, J. Sun, F. Juhn, K. Brennan, K. Iwanik, A. Maillet, J. Buell, E. White, M. Zhao, S. Balasubramanian, S. Terzic, T. Richards, V. Banning, L. Garcia, K. Mahoney, Z. Zwirko, A. Donahue, H. Beltran, J. M. Mosquera, M. A. Rubin, S. Dogan, C. V. Hedvat, M. F. Berger, L. Pusztai, M. Lechner, C. Boshoff, M. Jarosz, C. Vietz, A. Parker, V. A. Miller, J. S. Ross, J. Curran, M. T. Cronin, P. J. Stephens, D. Lipson, and R. Yelensky. Development and validation of a clinical cancer genomic profiling test based on massively parallel dna sequencing. *Nat Biotechnol*, 31(11):1023–31, 2013. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). 10.1038/nbt.2696. URL https://www.ncbi.nlm.nih.gov/pubmed/24142049.

D. S. Ross, A. Zehir, D. T. Cheng, R. Benayed, K. Nafa, J. F. Hechtman, Y. Y. Janjigian, B. Weigelt, P. Razavi, D. M. Hyman, J. Baselga, M. F. Berger, M. Ladanyi, and M. E. Arcila. Next-generation assessment of human epidermal growth factor receptor 2 (erbb2) amplification status: Clinical validation in the context of a hybrid capture-based, comprehensive solid tumor genomic profiling assay. *J Mol Diagn*, 19(2):244–254, 2017. ISSN 1943-7811 (Electronic) 1525-1578 (Linking). 10.1016/j.jmoldx.2016.09.010. URL https://www.ncbi.nlm.nih.gov/pubmed/28027945.

R. Bose, S. M. Kavuri, A. C. Searleman, W. Shen, D. Shen, D. C. Koboldt, J. Monsey, N. Goel, A. B. Aronson, S. Li, C. X. Ma, L. Ding, E. R. Mardis, and M. J. Ellis. Activating her2 mutations in her2 gene amplification negative breast cancer. *Cancer Discov*, 3(2):224–37, 2013. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). 10.1158/2159-8290.CD-12-0349. URL http://www.ncbi.nlm.nih.gov/pubmed/23220880http://cancerdiscovery.aacrjournals.org/content/3/2/224.full.pdf.

A. B. Hanker, M. R. Brewer, J. H. Sheehan, J. P. Koch, G. R. Sliwoski, R. Nagy, R. Lanman, M. F. Berger, D. M. Hyman, D. B. Solit, J. He, V. Miller, Jr. Cutler, R. E., A. S. Lalani, D. Cross, C. M. Lovly, J. Meiler, and C. L. Arteaga. An acquired her2(t798i) gatekeeper mutation induces resistance to neratinib in a patient with

her2 mutant-driven breast cancer. *Cancer Discov*, 7(6):575–585, 2017. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). 10.1158/2159-8290.CD-16-1431. URL https://www.ncbi.nlm.nih.gov/pubmed/28274957.

D. M. Hyman, S. A. Piha-Paul, H. Won, J. Rodon, C. Saura, G. I. Shapiro, D. Juric, D. I. Quinn, V. Moreno, B. Doger, I. A. Mayer, V. Boni, E. Calvo, S. Loi, A. C. Lockhart, J. P. Erinjeri, M. Scaltriti, G. A. Ulaner, J. Patel, J. Tang, H. Beer, S. D. Selcuklu, A. J. Hanrahan, N. Bouvier, M. Melcer, R. Murali, A. M. Schram, L. M. Smyth, K. Jhaveri, B. T. Li, A. Drilon, J. J. Harding, G. Iyer, B. S. Taylor, M. F. Berger, Jr. Cutler, R. E., F. Xu, A. Butturini, L. D. Eli, G. Mann, C. Farrell, A. S. Lalani, R. P. Bryce, C. L. Arteaga, F. Meric-Bernstam, J. Baselga, and D. B. Solit. Her kinase inhibition in patients with her2- and her3-mutant cancers. *Nature*, 554(7691):189–194, 2018. ISSN 0028-0836 (Print) 0028-0836. 10.1038/nature25475.

G. Deniziaut, J. C. Tille, F. C. Bidard, S. Vacher, A. Schnitzler, W. Chemlali, L. Tremoulet, L. Fuhrmann, P. Cottu, R. Rouzier, I. Bieche, and A. Vincent-Salomon. Erbb2 mutations associated with solid variant of high-grade invasive lobular breast carcinomas. *Oncotarget*, 7(45):73337–73346, 2016. ISSN 1949-2553 (Electronic) 1949-2553 (Linking). 10.18632/oncotarget.11819. URL https://www.ncbi.nlm.nih.gov/pubmed/27602491.

C. Desmedt, G. Zoppoli, G. Gundem, G. Pruneri, D. Larsimont, M. Fornili, D. Fumagalli, D. Brown, F. Rothe, D. Vincent, N. Kheddoumi, G. Rouas, S. Majjaj, S. Brohee, P. Van Loo, P. Maisonneuve, R. Salgado, T. Van Brussel, D. Lambrechts, R. Bose, O. Metzger, C. Galant, F. Bertucci, M. Piccart-Gebhart, G. Viale, E. Biganzoli, P. J. Campbell, and C. Sotiriou. Genomic characterization of primary invasive lobular breast cancer. *J Clin Oncol*, 34(16):1872–81, 2016. ISSN 1527-7755 (Electronic) 0732-183X (Linking). 10.1200/JCO.2015.64.0334. URL https://www.ncbi.nlm.nih.gov/pubmed/26926684.

Z. Ping, G. P. Siegal, S. Harada, I. E. Eltoum, M. Youssef, T. Shen, J. He, Y. Huang, D. Chen, Y. Li, K. I. Bland, H. R. Chang, and D. Shen. Erbb2 mutation is associated with a worse prognosis in patients with cdh1 altered invasive lobular cancer of the breast. *Oncotarget*, 7(49):80655–80663, 2016. ISSN 1949-2553. 10.18632/oncotarget.13019.

J. S. Ross, K. Wang, C. E. Sheehan, A. B. Boguniewicz, G. Otto, S. R. Downing, J. Sun, J. He, J. A. Curran, S. Ali, R. Yelensky, D. Lipson, G. Palmer, V. A. Miller, and P. J. Stephens. Relapsed classic e-cadherin (cdh1)-mutated invasive lobular breast cancer shows a high frequency of her2 (erbb2) gene mutations. *Clin Cancer Res*, 19(10):2668–76, 2013. ISSN 1078-0432 (Print) 1078-0432 (Linking). 10.1158/1078-0432.CCR-13-0295. URL https://www.ncbi.nlm.nih.gov/pubmed/23575477.

S. Kurozumi, M. Alsaleem, C. J. Monteiro, K. Bhardwaj, S. E. P. Joosten, T. Fujii, K. Shirabe, A. R. Green, I. O. Ellis, E. A. Rakha, N. P. Mongan, D. M. Heery, W. Zwart, S. Oesterreich, and S. J. Johnston. Targetable erbb2 mutation status is an independent marker of adverse prognosis in estrogen receptor positive, erbb2 non-amplified primary lobular breast carcinoma: a retrospective in silico analysis of public datasets. *Breast Cancer Res*, 22(1):85, 2020. ISSN 1465-542X (Electronic) 1465-5411 (Linking). 10.1186/s13058-020-01324-4. URL https://www.ncbi.nlm.nih.gov/pubmed/32782013.

T. Wang, Y. Xu, S. Sheng, H. Yuan, T. Ouyang, J. Li, T. Wang, Z. Fan, T. Fan, B. Lin, and Y. Xie. Her2 somatic mutations are associated with poor survival in her2-negative breast cancers. *Cancer Sci*, 108(4):671–677, 2017. ISSN 1349-7006 (Electronic) 1347-9032 (Linking). 10.1111/cas.13182. URL https://www.ncbi.nlm.nih.gov/pubmed/28164408.

S. Croessmann, L. Formisano, L. N. Kinch, P. I. Gonzalez-Ericsson, D. R. Sudhan, R. J. Nagy, A. Mathew, E. H. Bernicker, M. Cristofanilli, J. He, Jr. Cutler, R. E., A. S. Lalani, V. A. Miller, R. B. Lanman, N. V. Grishin, and C. L. Arteaga. Combined blockade of activating erbb2 mutations and er results in synthetic lethality of er+/her2 mutant breast cancer. *Clin Cancer Res*, 25(1):277–289, 2019. ISSN 1078-0432 (Print) 1078-0432 (Linking). 10.1158/1078-0432.CCR-18-1544. URL https://www.ncbi.nlm.nih.gov/pubmed/30314968.

U. Nayar, O. Cohen, C. Kapstad, M. S. Cuoco, A. G. Waks, S. A. Wander, C. Painter, S. Freeman, N. S. Persky, L. Marini, K. Helvie, N. Oliver, O. Rozenblatt-Rosen, C. X. Ma, A. Regev, E. P. Winer, N. U. Lin, and N. Wagle. Acquired her2 mutations in er(+) metastatic breast cancer confer resistance to estrogen receptor-directed therapies. *Nat Genet*, 51(2):207–216, 2019. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). 10.1038/s41588-018-0287-5. URL https://www.ncbi.nlm.nih.gov/pubmed/30531871.

P. Razavi, M. T. Chang, G. Xu, C. Bandlamudi, D. S. Ross, N. Vasan, Y. Cai, C. M. Bielski, M. T. A. Donoghue, P. Jonsson, A. Penson, R. Shen, F. Pareja, R. Kundra, S. Middha, M. L. Cheng, A. Zehir,

C. Kandoth, R. Patel, K. Huberman, L. M. Smyth, K. Jhaveri, S. Modi, T. A. Traina, C. Dang, W. Zhang, B. Weigelt, B. T. Li, M. Ladanyi, D. M. Hyman, N. Schultz, M. E. Robson, C. Hudis, E. Brogi, A. Viale, L. Norton, M. N. Dickler, M. F. Berger, C. A. Iacobuzio-Donahue, S. Chandarlapaty, M. Scaltriti, J. S. Reis-Filho, D. B. Solit, B. S. Taylor, and J. Baselga. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell*, 34(3):427–438 e6, 2018. ISSN 1878-3686 (Electronic) 1535-6108 (Linking). 10.1016/j.ccell.2018.08.008. URL https://www.ncbi.nlm.nih.gov/pubmed/30205045.

E. Cocco, F. Javier Carmona, P. Razavi, H. H. Won, Y. Cai, V. Rossi, C. Chan, J. Cownie, J. Soong, E. Toska, S. G. Shifman, I. Sarotto, P. Savas, M. J. Wick, K. P. Papadopoulos, A. Moriarty, Jr. Cutler, R. E., F. Avogadri-Connors, A. S. Lalani, R. P. Bryce, S. Chandarlapaty, D. M. Hyman, D. B. Solit, V. Boni, S. Loi, J. Baselga, M. F. Berger, F. Montemurro, and M. Scaltriti. Neratinib is effective in breast tumors bearing both amplification and mutation of erbb2 (her2). *Sci Signal*, 11(551), 2018. ISSN 1945-0877 (Print) 1945-0877. 10.1126/scisignal.aat9773.

X. Xu, C. De Angelis, K. A. Burke, A. Nardone, H. Hu, L. Qin, J. Veeraraghavan, V. Sethunath, L. M. Heiser, N. Wang, C. K. Y. Ng, E. S. Chen, A. Renwick, T. Wang, S. Nanda, M. Shea, T. Mitchell, M. Rajendran, I. Waters, D. J. Zabransky, K. L. Scott, C. Gutierrez, C. Nagi, F. C. Geyer, G. C. Chamness, B. H. Park, C. A. Shaw, S. G. Hilsenbeck, M. F. Rimawi, J. W. Gray, B. Weigelt, J. S. Reis-Filho, C. K. Osborne, and R. Schiff. Her2 reactivation through acquisition of the her2 l755s mutation as a mechanism of acquired resistance to her2-targeted therapy in her2(+) breast cancer. *Clin Cancer Res*, 23(17):5123–5134, 2017. ISSN 1078-0432 (Print) 1078-0432 (Linking). 10.1158/1078-0432.CCR-16-2191. URL https://www.ncbi.nlm.nih.gov/pubmed/28487443.

D. M. Hyman, S. Piha-Paul, J. Rodon, C. Saura, G. I. Shapiro, D. I. Quinn, and V.a Moreno. Neratinib in her2- or her3-mutant solid tumors: Summit, a global, multi-histology, open-label, phase 2 'basket' study, April 2017 2017.

C. L. Arteaga and J. A. Engelman. Erbb receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. *Cancer Cell*, 25(3):282–303, 2014. ISSN 1878-3686 (Electronic) 1535-6108 (Linking). 10.1016/j.ccr.2014.02.025. URL http://www.ncbi.nlm.nih.gov/pubmed/24651011.

C. Wallasch, F. U. Weiss, G. Niederfellner, B. Jallal, W. Issing, and A. Ullrich. Heregulin-dependent regulation of her2/neu oncogenic signaling by heterodimerization with her3. *Embo j*, 14(17):4267–75, 1995. ISSN 0261-4189 (Print) 0261-4189.

D. J. Zabransky, C. L. Yankaskas, R. L. Cochran, H. Y. Wong, S. Croessmann, D. Chu, S. M. Kavuri, M. Red Brewer, D. M. Rosen, W. B. Dalton, A. Cimino-Mathews, K. Cravero, B. Button, K. Kyker-Snowman, J. Cidado, B. Erlanger, H. A. Parsons, K. M. Manto, R. Bose, J. Lauring, C. L. Arteaga, K. Konstantopoulos, and B. H. Park. Her2 missense mutations have distinct effects on oncogenic signaling and migration. *Proc Natl Acad Sci U S A*, 112(45):E6205–14, 2015. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). 10.1073/pnas.1516853112. URL https://www.ncbi.nlm.nih.gov/pubmed/26508629.

Network Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). 10.1038/nature11412. URL https://www.ncbi.nlm.nih.gov/pubmed/23000897.

B. S. Jaiswal, N. M. Kljavin, E. W. Stawiski, E. Chan, C. Parikh, S. Durinck, S. Chaudhuri, K. Pujara, J. Guillory, K. A. Edgar, V. Janakiraman, R. P. Scholz, K. K. Bowman, M. Lorenzo, H. Li, J. Wu, W. Yuan, B. A. Peters, Z. Kan, J. Stinson, M. Mak, Z. Modrusan, C. Eigenbrot, R. Firestein, H. M. Stern, K. Rajalingam, G. Schaefer, M. A. Merchant, M. X. Sliwkowski, F. J. de Sauvage, and S. Seshagiri. Oncogenic erbb3 mutations in human cancers. *Cancer Cell*, 23(5):603–17, 2013. ISSN 1878-3686 (Electronic) 1535-6108 (Linking). 10.1016/j.ccr.2013.04.012. URL https://www.ncbi.nlm.nih.gov/pubmed/23680147.

B. Choi, M. Cha, G. S. Eun, D. H. Lee, S. Lee, M. Ehsan, P. S. Chae, W. D. Heo, Y. Park, and T. Y. Yoon. Single-molecule functional anatomy of endogenous her2-her3 heterodimers. *Elife*, 9, 2020. ISSN 2050-084x. 10.7554/eLife.53934.

T. Holbro, R. R. Beerli, F. Maurer, M. Koziczak, 3rd Barbas, C. F., and N. E. Hynes. The erbb2/erbb3 heterodimer functions as an oncogenic unit: Erbb2 requires erbb3 to drive breast tumor cell proliferation. *Proc Natl Acad Sci U S A*, 100(15):8933–8, 2003. ISSN 0027-8424 (Print) 0027-8424 (Linking). 10.1073/pnas.1537685100. URL https://www.ncbi.nlm.nih.gov/pubmed/12853564.

Y. Yarden and M. X. Sliwkowski. Untangling the erbb signalling network. *Nat Rev Mol Cell Biol*, 2(2): 127–37, 2001. ISSN 1471-0072 (Print) 1471-0072 (Linking). 10.1038/35052073. URL https://www.ncbi.nlm.nih.gov/pubmed/11252954.

L. M. Smyth, S. A. Piha-Paul, H. H. Won, A. M. Schram, C. Saura, S. Loi, J. Lu, G. I. Shapiro, D. Juric, I. A. Mayer, C. L. Arteaga, M. I. de la Fuente, A. M. Brufksy, I. Spanggaard, M. Mau-Sorensen, M. Arnedos, V. Moreno, V. Boni, J. Sohn, L. S. Schwartzberg, X. Gonzalez-Farre, A. Cervantes, F. C. Bidard, A. N. Gorelick, R. B. Lanman, R. J. Nagy, G. A. Ulaner, S. Chandarlapaty, K. Jhaveri, E. I. Gavrila, C. Zimel, S. D. Selcuklu, M. Melcer, A. Samoila, Y. Cai, M. Scaltriti, G. Mann, F. Xu, L. D. Eli, M. Dujka, A. S. Lalani, R. Bryce, J. Baselga, B. S. Taylor, D. B. Solit, F. Meric-Bernstam, and D. M. Hyman. Efficacy and determinants of response to her kinase inhibition in her2-mutant metastatic breast cancer. *Cancer Discov*, 10 (2):198–213, 2020. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). 10.1158/2159-8290.CD-19-0966. URL https://www.ncbi.nlm.nih.gov/pubmed/31806627.

Timothy S. Collier, Karthikeyan Diraviyam, John Monsey, Wei Shen, David Sept, and Ron Bose. Carboxyl group footprinting mass spectrometry and molecular dynamics identify key interactions in the her2-her3 receptor tyrosine kinase interface. *The Journal of biological chemistry*, 288(35):25254–25264, 2013. ISSN 1083-351X 0021-9258. 10.1074/jbc.M113.474882. URL https://pubmed.ncbi.nlm.nih.gov/23843458https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3757188/.

Peter Littlefield, Lijun Liu, Venkatesh Mysore, Yibing Shan, David E. Shaw, and Natalia Jura. Structural analysis of the egfr/her3 heterodimer reveals the molecular basis for activating her3 mutations. *Science Signaling*, 7(354):ra114, 2014. 10.1126/scisignal.2005786. URL http://stke.sciencemag.org/content/7/354/ra114.abstract.

J. P. Robichaux, Y. Y. Elamin, R. S. K. Vijayan, M. B. Nilsson, L. Hu, J. He, F. Zhang, M. Pisegna, A. Poteete, H. Sun, S. Li, T. Chen, H. Han, M. V. Negrao, J. R. Ahnert, L. Diao, J. Wang, X. Le, F. Meric-Bernstam, M. Routbort, B. Roeck, Z. Yang, V. M. Raymond, R. B. Lanman, G. M. Frampton, V. A. Miller, A. B. Schrock, L. A. Albacker, K. K. Wong, J. B. Cross, and J. V. Heymach. Pan-cancer landscape and analysis of erbb2 mutations identifies poziotinib as a clinically active inhibitor and enhancer of t-dm1 activity. *Cancer Cell*, 36(4):444–457 e7, 2019. ISSN 1878-3686 (Electronic) 1535-6108 (Linking). 10.1016/j.ccell.2019.09.001. URL https://www.ncbi.nlm.nih.gov/pubmed/31588020.

D. R. Sudhan, A. Guerrero-Zotano, H. Won, P. González Ericsson, A. Servetto, M. Huerta-Rosario, D. Ye, K. M. Lee, L. Formisano, Y. Guo, Q. Liu, L. N. Kinch, M. Red Brewer, T. Dugger, J. Koch, M. J. Wick, Jr. Cutler, R. E., A. S. Lalani, R. Bryce, A. Auerbach, A. B. Hanker, and C. L. Arteaga. Hyperactivation of torc1 drives resistance to the pan-her tyrosine kinase inhibitor neratinib in her2-mutant cancers. *Cancer Cell*, 37 (2):183–199.e5, 2020. ISSN 1535-6108. 10.1016/j.ccell.2019.12.013.

D. B. Agus, R. W. Akita, W. D. Fox, G. D. Lewis, B. Higgins, P. I. Pisacane, J. A. Lofgren, C. Tindell, D. P. Evans, K. Maiese, H. I. Scher, and M. X. Sliwkowski. Targeting ligand-activated erbb2 signaling inhibits breast and prostate tumor growth. *Cancer Cell*, 2(2):127–37, 2002. ISSN 1535-6108 (Print) 1535-6108 (Linking). 10.1016/s1535-6108(02)00097-1. URL https://www.ncbi.nlm.nih.gov/pubmed/12204533.

T. T. Junttila, R. W. Akita, K. Parsons, C. Fields, G. D. Lewis Phillips, L. S. Friedman, D. Sampath, and M. X. Sliwkowski. Ligand-independent her2/her3/pi3k complex is disrupted by trastuzumab and is effectively inhibited by the pi3k inhibitor gdc-0941. *Cancer Cell*, 15(5):429–40, 2009. ISSN 1878-3686 (Electronic) 1535-6108 (Linking). 10.1016/j.ccr.2009.03.020. URL https://www.ncbi.nlm.nih.gov/pubmed/19411071.

H. J. Jacobsen, T. T. Poulsen, A. Dahlman, I. Kjaer, K. Koefoed, J. W. Sen, D. Weilguny, B. Bjerregaard, C. R. Andersen, I. D. Horak, M. W. Pedersen, M. Kragh, and J. Lantto. Pan-her, an antibody mixture simultaneously targeting egfr, her2, and her3, effectively overcomes tumor heterogeneity and plasticity. *Clin Cancer Res*, 21(18):4110–22, 2015. ISSN 1078-0432 (Print) 1078-0432 (Linking). 10.1158/1078-0432.CCR-14-3312. URL https://www.ncbi.nlm.nih.gov/pubmed/25908781.

C. X. Ma, R. Bose, F. Gao, R. A. Freedman, M. L. Telli, G. Kimmick, E. Winer, M. Naughton, M. P. Goetz, C. Russell, D. Tripathy, M. Cobleigh, A. Forero, T. J. Pluard, C. Anders, P. A. Niravath, S. Thomas, J. Anderson, C. Bumb, K. C. Banks, R. B. Lanman, R. Bryce, A. S. Lalani, J. Pfeifer, D. F. Hayes, M. Pegram, K. Blackwell, P. L. Bedard, H. Al-Kateb, and M. J. C. Ellis. Neratinib efficacy and circulating tumor dna

detection of her2 mutations in her2 nonamplified metastatic breast cancer. *Clin Cancer Res*, 23(19):5687–5695, 2017. ISSN 1078-0432 (Print) 1078-0432 (Linking). 10.1158/1078-0432.CCR-17-0900. URL https://www.ncbi.nlm.nih.gov/pubmed/28679771.

Jiayao Li, Qian Xiao, Yi Bao, Wenyu Wang, Jianyuan Goh, Panpan Wang, and Qiang Yu. Her2-l755s mutation induces hyperactive mapk and pi3k-mtor signaling, leading to resistance to her2 tyrosine kinase inhibitor treatment. *Cell cycle (Georgetown, Tex.)*, 18(13):1513–1522, 2019. ISSN 1551-4005 1538-4101. 10.1080/15384101.2019.1624113. URL https://pubmed.ncbi.nlm.nih.gov/31135266https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6592242/.

A. N. Gorelick, F. J. Sánchez-Rivera, Y. Cai, C. M. Bielski, E. Biederstedt, P. Jonsson, A. L. Richards, N. Vasan, A. V. Penson, N. D. Friedman, Y. J. Ho, T. Baslan, C. Bandlamudi, M. Scaltriti, N. Schultz, S. W. Lowe, E. Reznik, and B. S. Taylor. Phase and context shape the function of composite oncogenic mutations. *Nature*, 582(7810):100–103, 2020. ISSN 0028-0836 (Print) 0028-0836. 10.1038/s41586-020-2315-8.

Y. Saito, J. Koya, M. Araki, Y. Kogure, S. Shingaki, M. Tabata, M. B. McClure, K. Yoshifuji, S. Matsumoto, Y. Isaka, H. Tanaka, T. Kanai, S. Miyano, Y. Shiraishi, Y. Okuno, and K. Kataoka. Landscape and function of multiple mutations within individual oncogenes. *Nature*, 582(7810):95–99, 2020. ISSN 0028-0836. 10.1038/s41586-020-2175-2.

N. Vasan, P. Razavi, J. L. Johnson, H. Shao, H. Shah, A. Antoine, E. Ladewig, A. Gorelick, T. Y. Lin, E. Toska, G. Xu, A. Kazmi, M. T. Chang, B. S. Taylor, M. N. Dickler, K. Jhaveri, S. Chandarlapaty, R. Rabadan, E. Reznik, M. L. Smith, R. Sebra, F. Schimmoller, T. R. Wilson, L. S. Friedman, L. C. Cantley, M. Scaltriti, and J. Baselga. Double pik3ca mutations in cis increase oncogenicity and sensitivity to pi3k inhibitors. *Science*, 366(6466):714–723, 2019. ISSN 0036-8075 (Print) 0036-8075. 10.1126/science.aaw9032.

T. Smirnova, Z. N. Zhou, R. J. Flinn, J. Wyckoff, P. J. Boimel, M. Pozzuto, S. J. Coniglio, J. M. Backer, A. R. Bresnick, J. S. Condeelis, N. E. Hynes, and J. E. Segall. Phosphoinositide 3-kinase signaling is critical for erbb3-driven breast cancer cell motility and metastasis. *Oncogene*, 31(6):706–15, 2012. ISSN 1476-5594 (Electronic) 0950-9232 (Linking). 10.1038/onc.2011.275. URL https://www.ncbi.nlm.nih.gov/pubmed/21725367.

C. Xue, F. Liang, R. Mahmood, M. Vuolo, J. Wyckoff, H. Qian, K. L. Tsai, M. Kim, J. Locker, Z. Y. Zhang, and J. E. Segall. Erbb3-dependent motility and intravasation in breast cancer metastasis. *Cancer Res*, 66(3):1418–26, 2006. ISSN 0008-5472 (Print) 0008-5472 (Linking). 10.1158/0008-5472.CAN-05-0550. URL https://www.ncbi.nlm.nih.gov/pubmed/16452197.

A. Chakrabarty, B. N. Rexer, S. E. Wang, R. S. Cook, J. A. Engelman, and C. L. Arteaga. H1047r phosphatidylinositol 3-kinase mutant enhances her2-mediated transformation by heregulin production and activation of her3. *Oncogene*, 29(37):5193–203, 2010. ISSN 1476-5594 (Electronic) 0950-9232 (Linking). 10.1038/onc.2010.257. URL https://www.ncbi.nlm.nih.gov/pubmed/20581867.

A. B. Hanker, A. D. Pfefferle, J. M. Balko, M. G. Kuba, C. D. Young, V. Sanchez, C. R. Sutton, H. Cheng, C. M. Perou, J. J. Zhao, R. S. Cook, and C. L. Arteaga. Mutant pik3ca accelerates her2-driven transgenic mammary tumors and induces resistance to combinations of anti-her2 therapies. *Proc Natl Acad Sci U S A*, 110(35):14372–7, 2013. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). 10.1073/pnas.1303204110. URL https://www.ncbi.nlm.nih.gov/pubmed/23940356.

B. N. Rexer, S. Chanthaphaychith, K. Dahlman, and C. L. Arteaga. Direct inhibition of pi3k in combination with dual her2 inhibitors is required for optimal antitumor activity in her2+ breast cancer cells. *Breast Cancer Res*, 16(1):R9, 2014. ISSN 1465-542X (Electronic) 1465-5411 (Linking). 10.1186/bcr3601. URL https://www.ncbi.nlm.nih.gov/pubmed/24451154.

S. Jain, A. N. Shah, C. A. Santa-Maria, K. Siziopikou, A. Rademaker, I. Helenowski, M. Cristofanilli, and W. J. Gradishar. Phase i study of alpelisib (byl-719) and trastuzumab emtansine (t-dm1) in her2-positive metastatic breast cancer (mbc) after trastuzumab and taxane therapy. *Breast Cancer Res Treat*, 171(2):371–381, 2018. ISSN 0167-6806. 10.1007/s10549-018-4792-0.

E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cbio cancer genomics

portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5):401–4, 2012. ISSN 2159-8290 (Electronic) 2159-8274 (Linking). 10.1158/2159-8290.CD-12-0095. URL https://www.ncbi.nlm.nih.gov/pubmed/22588877.

B. J. Bender, 3rd Cisneros, A., A. M. Duran, J. A. Finn, D. Fu, A. D. Lokits, B. K. Mueller, A. K. Sangha, M. F. Sauer, A. M. Sevy, G. Sliwoski, J. H. Sheehan, F. DiMaio, J. Meiler, and R. Moretti. Protocols for molecular modeling with rosetta3 and rosettascripts. *Biochemistry*, 55(34):4748–63, 2016. ISSN 1520-4995 (Electronic) 0006-2960 (Linking). 10.1021/acs.biochem.6b00444. URL https://www.ncbi.nlm.nih.gov/pubmed/27490953http://pubs.acs.org/doi/pdfplus/10.1021/acs.biochem.6b00444.

J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P. S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmstrom, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. O'Conchuir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D. A. Silva, S. Smith, Y. Song, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nat Methods*, 17(7): 665–680, 2020. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). 10.1038/s41592-020-0848-2. URL https://www.ncbi.nlm.nih.gov/pubmed/32483333.

Steven A. Combs, Samuel L. DeLuca, Stephanie H. DeLuca, Gordon H. Lemmon, David P. Nannemann, Elizabeth D. Nguyen, Jordan R. Willis, Jonathan H. Sheehan, and Jens Meiler. Small-molecule ligand docking into comparative models with rosetta. *Nature Protocols*, 8(7):1277–1298, 2013. ISSN 1750-2799. 10.1038/nprot.2013.074. URL https://doi.org/10.1038/nprot.2013.074.

J. Meiler and D. Baker. Rosettaligand: protein-small molecule docking with full side-chain flexibility. *Proteins*, 65(3):538–48, 2006. ISSN 1097-0134 (Electronic) 0887-3585 (Linking). 10.1002/prot.21086. URL http://www.ncbi.nlm.nih.gov/pubmed/16972285.

James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015. ISSN 1549-9618. 10.1021/acs.jctc.5b00255. URL https://doi.org/10.1021/acs.jctc.5b00255.

Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, and Peter A. Kollman. Application of resp charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society*, 115(21):9620–9631, 1993. ISSN 0002-7863. 10.1021/ja00074a030. URL https://doi.org/10.1021/ja00074a030.

K. Aertgeerts, R. Skene, J. Yano, B. C. Sang, H. Zou, G. Snell, A. Jennings, K. Iwamoto, N. Habuka, A. Hirokawa, T. Ishikawa, T. Tanaka, H. Miki, Y. Ohta, and S. Sogabe. Structural analysis of the mechanism of inhibition and allosteric activation of the kinase domain of her2 protein. *J Biol Chem*, 286(21):18756–65, 2011. ISSN 1083-351X (Electronic) 0021-9258 (Linking). 10.1074/jbc.M110.206193. URL http://www.ncbi.nlm.nih.gov/pubmed/21454582.

Julia Koehler Leman and Richard Bonneau. A novel domain assembly routine for creating full-length models of membrane proteins from known domain structures. *Biochemistry*, 57(13):1939–1944, 2018. ISSN 0006-2960. 10.1021/acs.biochem.7b00995. URL https://doi.org/10.1021/acs.biochem.7b00995.

Jeffrey Mendenhall, Benjamin P. Brown, Sandeepkumar Kothiwale, and Jens Meiler. Bcl::conf: Improved open-source knowledge-based conformation sampling using the crystallography open database. *Journal of Chemical Information and Modeling*, 2020. ISSN 1549-9596. 10.1021/acs.jcim.0c01140. URL https://doi.org/10.1021/acs.jcim.0c01140.

S. DeLuca, K. Khar, and J. Meiler. Fully flexible docking of medium sized ligand libraries with rosettaligand. *PLoS One*, 10(7):e0132508, 2015. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). 10.1371/journal.pone.0132508. URL http://www.ncbi.nlm.nih.gov/pubmed/26207742http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4514752/pdf/pone.0132508.pdf.

Carlos Vega and Jose L. F. Abascal. Simulating water with rigid non-polarizable models: a general perspective. *Physical Chemistry Chemical Physics*, 13(44):19663–19688, 2011. ISSN 1463-9076. 10.1039/C1CP22168J. URL http://dx.doi.org/10.1039/C1CP22168J.

Jörg Weiser, Peter S. Shenkin, and W. Clark Still. Approximate atomic surfaces from linear combinations of pairwise overlaps (lcpo). *Journal of Computational Chemistry*, 20(2):217–230, 1999. ISSN 0192-8651. 10.1002/(SICI)1096-987X(19990130)20:2¡217::AID-JCC4¿3.0.CO;2-A. URL https://doi.org/10.1002/(SICI)1096-987X(19990130)20:2⟨217::AID-JCC4⟩3.0.CO;2-A.

Brandon Frenz, Steven M. Lewis, Indigo King, Frank DiMaio, Hahnbeom Park, and Yifan Song. Prediction of protein mutational free energy: Benchmark and sampling improvements increase classification accuracy. *Frontiers in Bioengineering and Biotechnology*, 8(1175), 2020. ISSN 2296-4185. 10.3389/fbioe.2020.558247. URL https://www.frontiersin.org/article/10.3389/fbioe.2020.558247.

Rodrigo Prado Martins, Sarah Findakly, Chrysoula Daskalogianni, Marie-Paule Teulade-Fichou, Marc Blondel, and Robin Fåhraeus. In cellulo protein-mrna interaction assay to determine the action of g-quadruplex-binding molecules. *Molecules*, 23(12), 2018. ISSN 1420-3049. 10.3390/molecules23123124.

J. Debnath, S. K. Muthuswamy, and J. S. Brugge. Morphogenesis and oncogenesis of mcf-10a mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods*, 30(3):256–68, 2003. ISSN 1046-2023 (Print) 1046-2023 (Linking). 10.1016/s1046-2023(03)00032-x. URL https://www.ncbi.nlm.nih.gov/pubmed/12798140.

Gerhard Wolber, Thomas Seidel, Fabian Bendix, and Thierry Langer. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today*, 13(1):23–29, 2008. ISSN 1359-6446. 10.1016/j.drudis.2007.09.007. URL http://www.sciencedirect.com/science/article/pii/S1359644607003996https://www.sciencedirect.com/science/article/pii/S1359644607003996?via%3Dihubhttps://www.sciencedirect.com/science/article/pii/S1359644607003996.

M. Vieth, J. D. Hirst, and 3rd Brooks, C. L. Do active site conformations of small ligands correspond to low free-energy solution structures? *J Comput Aided Mol Des*, 12(6):563–72, 1998. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9879504.

Emanuele Perola and Paul S. Charifson. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *Journal of Medicinal Chemistry*, 47(10):2499–2510, 2004. ISSN 0022-2623. 10.1021/jm030563w. URL https://doi.org/10.1021/jm030563w.

Ming-Hong Hao, Omar Haq, and Ingo Muegge. Torsion angle preference and energetics of small-molecule ligands bound to proteins. *Journal of Chemical Information and Modeling*, 47(6):2242–2252, 2007. ISSN 1549-9596. 10.1021/ci700189s. URL https://doi.org/10.1021/ci700189s.

Moon-Hyeong Seo, Jeongbin Park, Eunkyung Kim, Sungchul Hohng, and Hak-Sung Kim. Protein conformational dynamics dictate the binding affinity for a ligand. *Nature Communications*, 5:ncomms4724, 2014. ISSN 2041-1723. 10.1038/ncomms4724. URL https://www.nature.com/articles/ncomms4724.

Nicholas Greives and Huan-Xiang Zhou. Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. *Proceedings of the National Academy of Sciences*, 111(28):10197–10202, 2014. ISSN 0027-8424, 1091-6490. 10.1073/pnas.1407545111. URL http://www.pnas.org/content/111/28/10197http://www.ncbi.nlm.nih.gov/pubmed/24982141http://www.pnas.org/content/111/28/10197.full.

Sheng-Yong Yang. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*, 15(11):444–450, 2010. ISSN 1359-6446. 10.1016/j.drudis.2010.03.013. URL http://www.sciencedirect.com/science/article/pii/S135964461000111Xhttps://www.sciencedirect.com/science/article/pii/S135964461000111X#bib20.

Dimitrios Vlachakis, Paraskevas Fakourelis, Vasileios Megalooikonomou, Christos Makris, and Sophia Kossida. Drugon: a fully integrated pharmacophore modeling and structure optimization toolkit. *PeerJ*, 3:e725, 2015. ISSN 2167-8359. 10.7717/peerj.725. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4304849/.

Andrea R. Beccari, Carlo Cavazzoni, Claudia Beato, and Gabriele Costantino. Ligen: A high performance workflow for chemistry driven de novo design. *Journal of Chemical Information and Modeling*, 53(6): 1518–1527, 2013. ISSN 1549-9596. 10.1021/ci400078g. URL https://doi.org/10.1021/ci400078g.

Evan A. Hecker, Chaya Duraiswami, Tariq A. Andrea, and David J. Diller. Use of catalyst pharmacophore models for screening of large combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 42(5):1204–1211, 2002. ISSN 0095-2338. 10.1021/ci020368a. URL https://doi.org/10.1021/ci020368ahttps://pubs.acs.org/doi/full/10.1021/ci020368a.

Hugo Kubinyi. *Comparative Molecular Field Analysis (CoMFA)*, pages 1555–1574. Wiley-VCH Verlag GmbH, 2003. ISBN 978-3-527-61827-9. URL http://onlinelibrary.wiley.com/doi/10.1002/9783527618279.ch44d/summary.

A. E. Cleves and A. N. Jain. Quantitative surface field analysis: learning causal models to predict ligand binding affinity and pose. *J Comput Aided Mol Des*, 2018. ISSN 0920-654x. 10.1007/s10822-018-0126-x.

S. L. Chan. Molalign: an algorithm for aligning multiple small molecules. *J Comput Aided Mol Des*, 31(6): 523–546, 2017. ISSN 0920-654x. 10.1007/s10822-017-0023-8.

Ambrish Roy and Jeffrey Skolnick. Ligsift: an open-source tool for ligand structural alignment and virtual screening. *Bioinformatics*, 31(4):539–544, 2015. ISSN 1367-4803 1367-4811. 10.1093/bioinformatics/btu692. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4325547/.

R. D. Urniaz and K. Jozwiak. X-ray crystallographic structures as a source of ligand alignment in 3d-qsar. *J Chem Inf Model*, 53(6):1406–14, 2013. ISSN 1549-9596. 10.1021/ci400004e.

M. Thormann, A. Klamt, and K. Wichmann. Cosmosim3d: 3d-similarity and alignment based on cosmo polarization charge densities. *J Chem Inf Model*, 52(8):2149–56, 2012. ISSN 1549-9596. 10.1021/ci300205p.

G. M. Sastry, S. L. Dixon, and W. Sherman. Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. *J Chem Inf Model*, 51(10): 2455–66, 2011. ISSN 1549-9596. 10.1021/ci2002704.

P. Tosco, T. Balle, and F. Shiri. Open3dalign: an open-source software aimed at unsupervised ligand alignment. *J Comput Aided Mol Des*, 25(8):777–83, 2011. ISSN 0920-654x. 10.1007/s10822-011-9462-9.

O. Korb, P. Monecke, G. Hessler, T. Stutzle, and T. E. Exner. pharmacophore: multiple flexible ligand alignment based on ant colony optimization. *J Chem Inf Model*, 50(9):1669–81, 2010. ISSN 1549-9596. 10.1021/ci1000218.

A. Heifets and R. H. Lilien. Ligalign: flexible ligand-based active site alignment and analysis. *J Mol Graph Model*, 29(1):93–101, 2010. ISSN 1093-3263. 10.1016/j.jmgm.2010.05.005.

A. N. Jain. Surflex-dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J Comput Aided Mol Des*, 21(5):281–306, 2007. ISSN 0920-654X (Print) 0920-654x. 10.1007/s10822-007-9114-2.

N. J. Richmond, C. A. Abrams, P. R. Wolohan, E. Abrahamian, P. Willett, and R. D. Clark. Galahad: 1. pharmacophore identification by hypermolecular alignment of ligands in 3d. *J Comput Aided Mol Des*, 20 (9):567–87, 2006. ISSN 0920-654X (Print) 0920-654x. 10.1007/s10822-006-9082-y.

S. A. Wildman and G. M. Crippen. Evaluation of ligand overlap by atomic parameters. *J Chem Inf Comput Sci*, 41(2):446–50, 2001. ISSN 0095-2338 (Print) 0095-2338.

G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J. F. Truchon, and W. D. Cornell. Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model*, 47(4):1504–19, 2007. ISSN 1549-9596 (Print). 10.1021/ci700052x. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17591764http://pubs.acs.org/doi/pdfplus/10.1021/ci700052x.

G. J. Tawa, J. C. Baber, and C. Humblet. Computation of 3d queries for rocs based virtual screens. *J Comput Aided Mol Des*, 23(12):853–68, 2009. ISSN 1573-4951 (Electronic) 0920-654X (Linking). 10.1007/s10822-009-9302-3. URL http://www.ncbi.nlm.nih.gov/pubmed/19820902.

Ewgenij Proschak, Matthias Rupp, Swetlana Derksen, and Gisbert Schneider. Shapelets: Possibilities and limitations of shape-based virtual screening. *Journal of Computational Chemistry*, 29(1):108–114, 2007. ISSN 0192-8651. 10.1002/jcc.20770. URL https://doi.org/10.1002/jcc.20770.

Anu J. Tervo, Toni Rönkkö, Tommi H. Nyrönen, and Antti Poso. Brutus: Optimization of a grid-based similarity function for rigid-body molecular superposition. 1. alignment and virtual screening applications. *Journal of Medicinal Chemistry*, 48(12):4076–4086, 2005. ISSN 0022-2623. 10.1021/jm049123a. URL https://doi.org/10.1021/jm049123a.

Tim Cheeseright, Mark Mackey, Sally Rose, and Andy Vinter. Molecular field extrema as descriptors of biological activity: Definition and validation. *Journal of Chemical Information and Modeling*, 46(2):665–676, 2006. ISSN 1549-9596. 10.1021/ci050357s. URL https://doi.org/10.1021/ci050357s.

Woong-Hee Shin, Xiaolei Zhu, Mark Bures, and Daisuke Kihara. Three-dimensional compound comparison methods and their application in drug discovery. *Molecules*, 20(7):12841, 2015. ISSN 1420-3049. URL http://www.mdpi.com/1420-3049/20/7/12841.

Katherine M. Andrews and Richard D. Cramer. Toward general methods of targeted library design. topomer shape similarity searching with diverse structures as queries. *Journal of Medicinal Chemistry*, 43(9):1723–1740, 2000. URL http://pubs.acs.org/doi/pdfplus/10.1021/jm000003m.

S. Kothiwale, J. L. Mendenhall, and J. Meiler. Bcl::conf: small molecule conformational sampling using a knowledge based rotamer library. *J Cheminform*, 7:47, 2015. ISSN 1758-2946 (Electronic) 1758-2946 (Linking). 10.1186/s13321-015-0095-1. URL http://www.ncbi.nlm.nih.gov/pubmed/26473018http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4607025/pdf/13321_2015_Article_95.pdf.

Mikko J. Vainio, J. Santeri Puranen, and Mark S. Johnson. Shaep: Molecular overlay based on shape and electrostatic potential. *Journal of Chemical Information and Modeling*, 49(2):492–502, 2009. ISSN 1549-9596. 10.1021/ci800315d. URL https://doi.org/10.1021/ci800315d.

M. Karakas, N. Woetzel, R. Staritzbichler, N. Alexander, B. E. Weiner, and J. Meiler. Bcl::fold–de novo prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS One*, 7(11):e49240, 2012. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). 10.1371/journal.pone.0049240. URL http://www.ncbi.nlm.nih.gov/pubmed/23173050http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3500284/pdf/pone.0049240.pdf.

James J. Kuffner. Effective sampling and distance metrics for 3d rigid body path planning. *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, 4:3993–3998 Vol.4, 2004.

Karen J. Gregory, Elizabeth D. Nguyen, Chrysa Malosh, Jeffrey L. Mendenhall, Jessica Z. Zic, Brittney S. Bates, Meredith J. Noetzel, Emma F. Squire, Eric M. Turner, Jerri M. Rook, Kyle A. Emmitte, Shaun R. Stauffer, Craig W. Lindsley, Jens Meiler, and P. Jeffrey Conn. Identification of specific ligand–receptor interactions that govern binding and cooperativity of diverse modulators to a common metabotropic glutamate receptor 5 allosteric site. *ACS Chemical Neuroscience*, 5(4):282–295, 2014. 10.1021/cn400225x. URL https://doi.org/10.1021/cn400225x.

Ilenia Giangreco, David A. Cosgrove, and Martin J. Packer. An extensive and diverse set of molecular overlays for the validation of pharmacophore programs. *Journal of Chemical Information and Modeling*, 53(4):852–866, 2013. ISSN 1549-9596. 10.1021/ci400020a. URL https://doi.org/10.1021/ci400020a.

Qi Chen, Richard E. Higgs, and Michal Vieth. Geometric accuracy of three-dimensional molecular overlays. *Journal of Chemical Information and Modeling*, 46(5):1996–2002, 2006. ISSN 1549-9596. 10.1021/ci060134h. URL https://doi.org/10.1021/ci060134hhttps://pubs.acs.org/doi/full/10.1021/ci060134h.

Darwin Yu Fu and Jens Meiler. Rosettaligandensemble: A small-molecule ensemble-driven docking approach. *ACS Omega*, 3(4):3655–3664, 2018. ISSN 2470-1343. 10.1021/acsomega.7b02059. URL https://doi.org/10.1021/acsomega.7b02059.

G. Lemmon and J. Meiler. Rosetta ligand docking with flexible xml protocols. *Methods Mol Biol*, 819: 143–55, 2012. ISSN 1940-6029 (Electronic) 1064-3745 (Linking). 10.1007/978-1-61779-465-$0_1$0.*URL*.

K. W. Kaufmann and J. Meiler. Using rosettaligand for small molecule docking into comparative models. *PLoS One*, 7(12):e50769, 2012. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). 10.1371/journal.pone.0050769. URL http://www.ncbi.nlm.nih.gov/pubmed/23239984http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3519832/pdf/pone.0050769.pdf.

I. W. Davis and D. Baker. Rosettaligand docking with full ligand and receptor flexibility. *J Mol Biol*, 385 (2):381–92, 2009. ISSN 1089-8638 (Electronic) 0022-2836 (Linking). 10.1016/j.jmb.2008.11.010. URL http://www.ncbi.nlm.nih.gov/pubmed/19041878.

I. W. Davis, K. Raha, M. S. Head, and D. Baker. Blind docking of pharmaceutically relevant compounds using rosettaligand. *Protein Sci*, 18(9):1998–2002, 2009. ISSN 1469-896X (Electronic) 0961-8368 (Linking). 10.1002/pro.192. URL http://www.ncbi.nlm.nih.gov/pubmed/19554568.

W.L. DeLano. The pymol molecular graphics system, 2007. URL http://www.pymol.org.

J. Gasteiger and Mario Marsili. Iterative partial equalization of oribital electronegativity - a rapid access to atomic charges. *Tetrahedron Letters*, 36:3219–3228, 1980.

Kenneth J. Miller. Additivity methods in molecular polarizability. *Journal of the American Chemical Society*, 112(23):8533–8542, 1990. ISSN 0002-7863. 10.1021/ja00179a044. URL https://doi.org/10.1021/ja00179a044.

Linus Pauling. The nature of the chemical bond. iv. the energy of single bonds and the relative electronegativity of atoms. *Journal of the American Chemical Society*, 54(9):3570–3582, 1932. ISSN 0002-7863. 10.1021/ja01348a011. URL https://doi.org/10.1021/ja01348a011.

Paul Labute. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, 18 (4):464–477, 2000. ISSN 1093-3263. https://doi.org/10.1016/S1093-3263(00)00068-1. URL http://www.sciencedirect.com/science/article/pii/S1093326300000681.

Aromaticity detection in marvin. a. URL http://onlinelibrarystatic.wiley.com/marvin/help/sci/aromatization-doc.htmlfiles/1157/aromatization-doc.html.

Daylight theory: Smiles. b. URL http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

George E. Dahl. Multi-task neural networks for qsar predictions. *arXiv preprint arXiv:1406.1231*, 2014.

Jeffrey Mendenhall and Jens Meiler. Improving quantitative structure-activity relationship models using artificial neural networks trained with dropout. *Journal of computer-aided molecular design*, 30(2):177–189, 2016. ISSN 1573-4951 0920-654X. 10.1007/s10822-016-9895-2. URL https://www.ncbi.nlm.nih.gov/pubmed/26830599https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4798928/.

A. Hillebrecht and G. Klebe. Use of 3d qsar models for database screening: a feasibility study. *J. Chem. Inf. Model*, 48(2):384–396, 2008. URL http://pubs.acs.org/doi/pdfplus/10.1021/ci7002945.

J. Manchester and R. Czerminski. Samfa: Simplifying molecular description for 3d-qsar. *J Chem Inf Model*, 2008. ISSN 1549-9596 (Print). 10.1021/ci800009u. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18503264http://pubs.acs.org/doi/pdfplus/10.1021/ci800009u.

Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003. URL http://dx.doi.org/10.1021/ci034160ghttp://pubs.acs.org/doi/pdfplus/10.1021/ci034160g.

Jr. Lowe, E. W., A. Ferrebee, A. L. Rodriguez, P. J. Conn, and J. Meiler. 3d-qsar comfa study of benzoxazepine derivatives as mglur5 positive allosteric modulators. *Bioorg Med Chem Lett*, 20(19):5922–4, 2010. ISSN 1464-3405 (Electronic) 0960-894X (Linking). 10.1016/j.bmcl.2010.07.061. URL http://www.ncbi.nlm.nih.gov/pubmed/20732812.

R. Mueller, A. L. Rodriguez, E. S. Dawson, M. Butkiewicz, T. T. Nguyen, S. Oleszkiewicz, A. Bleckmann, C. D. Weaver, C. W. Lindsley, P. J. Conn, and J. Meiler. Identification of metabotropic glutamate receptor subtype 5 potentiators using virtual high-throughput screening. *ACS Chem Neurosci*, 1(4):288–305, 2010.

ISSN 1948-7193 (Electronic) 1948-7193 (Linking). 10.1021/cn9000389. URL http://www.ncbi.nlm.nih.gov/pubmed/20414370.

Annalen Bleckmann and Jens Meiler. Epothilones: Quantitative structure activity relations studied by support vector machines and artificial neural networks. *QSAR Combinatorial Science*, 22(7):722–728, 2003. ISSN 1611-020X 1611-0218. 10.1002/qsar.200330837. URL ⟨GotoISI⟩://WOS:000186347700003.

Edward W. Lowe, Mariusz Butkiewicz, Nils Woetzel, and Jens Meiler. Gpu-accelerated machine learning techniques enable qsar modeling of large hts data. In *IEEE*, page 3140320.

G. Sliwoski, E. W. Lowe, M. Butkiewicz, and J. Meiler. Bcl::emas–enantioselective molecular asymmetry descriptor for 3d-qsar. *Molecules*, 17(8):9971–89, 2012. ISSN 1420-3049 (Electronic) 1420-3049 (Linking). 10.3390/molecules17089971. URL http://www.ncbi.nlm.nih.gov/pubmed/22907158.

G. Sliwoski, J. Mendenhall, and J. Meiler. Autocorrelation descriptor improvements for qsar: 2da$_s$*ignand*3*da*$_s$*ign.J Comput Aid* $4951(Electronic)0920-654X(Linking).10.1007/s10822-015-9893-9.URL.$

Jeffrey Mendenhall, Benjamin P. Brown, Sandeepkumar Kothiwale, and Jens Meiler. Bcl::conf: Improved open-source knowledge-based conformation sampling using the crystallography open database. *Journal of Chemical Information and Modeling*, 61(1):189–201, 2021. ISSN 1549-9596. 10.1021/acs.jcim.0c01140. URL https://doi.org/10.1021/acs.jcim.0c01140. doi: 10.1021/acs.jcim.0c01140.

Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John Z. H. Zhang, and Tingjun Hou. End-point binding free energy calculation with mm/pbsa and mm/gbsa: Strategies and applications in drug design. *Chemical Reviews*, 119(16):9478–9508, 2019c. ISSN 0009-2665. 10.1021/acs.chemrev.9b00055. URL https://doi.org/10.1021/acs.chemrev.9b00055.

Huiyong Sun, Lili Duan, Fu Chen, Hui Liu, Zhe Wang, Peichen Pan, Feng Zhu, John Z. H. Zhang, and Tingjun Hou. Assessing the performance of mm/pbsa and mm/gbsa methods. 7. entropy effects on the performance of end-point binding free energy calculation approaches. *Physical chemistry chemical physics : PCCP*, 20(21):14450–14460, 2018. ISSN 1463-9084. 10.1039/c7cp07623a. URL https://pubmed.ncbi.nlm.nih.gov/29785435.

P. J. Ballester and J. B. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–75, 2010. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). btq112 [pii] 10.1093/bioinformatics/btq112. URL http://www.ncbi.nlm.nih.gov/pubmed/20236947http://bioinformatics.oxfordjournals.org/content/26/9/1169.full.pdf.

Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *Journal of computational chemistry*, 38(3):169–177, 2017. ISSN 1096-987X 0192-8651. 10.1002/jcc.24667. URL https://pubmed.ncbi.nlm.nih.gov/27859414https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5140681/.

Janaina Cruz Pereira, Ernesto Raúl Caffarena, and Cicero Nogueira dos Santos. Boosting docking-based virtual screening with deep learning. *Journal of Chemical Information and Modeling*, 56(12):2495–2506, 2016. ISSN 1549-9596. 10.1021/acs.jcim.6b00355. URL https://doi.org/10.1021/acs.jcim.6b00355.

Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein-ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017. ISSN 1549-960X 1549-9596. 10.1021/acs.jcim.6b00740. URL https://pubmed.ncbi.nlm.nih.gov/28368587https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5479431/.

José Jiménez, Miha Škalič, Gerard Martínez-Rosell, and Gianni De Fabritiis. Kdeep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296, 2018. ISSN 1549-9596. 10.1021/acs.jcim.7b00650. URL https://doi.org/10.1021/acs.jcim.7b00650.

Abraham Heifets Izhar Wallach, Michael Dzamba. Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *ArXiv preprint*, 2015.

N. Minovski, S. Zuperl, V. Drgan, and M. Novic. Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum euclidean distance space analysis: A case study. *Analytica Chimica Acta*, 759:28–42, 2013. ISSN 0003-2670. URL ⟨GotoISI⟩://WOS:

000313761300004http://ac.els-cdn.com/S000326701201642X/1-s2.0-S000326701201642X-main.pdf?_tid=3ffe46ea-8baf-11e
acdnat=1447602231_294955f3b06b5118580681387343cbf8.

R. P. Sheridan. Three useful dimensions for domain applicability in qsar models using random forest. *J Chem Inf Model*, 52(3):814–23, 2012. ISSN 1549-960X (Electronic) 1549-9596 (Linking). 10.1021/ci300004n. URL http://www.ncbi.nlm.nih.gov/pubmed/22385389.

Igor V. Tetko, Iurii Sushko, Anil Kumar Pandey, Hao Zhu, Alexander Tropsha, Ester Papa, Tomas Oberg, Roberto Todeschini, Denis Fourches, and Alexandre Varnek. Critical assessment of qsar models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.*, 48(9):1733–1746, 2008. ISSN 1549-9596. URL http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ci800151mhttp://pubs.acs.org/doi/pdfplus/10.1021/ci800151m.

Timon Schroeter, Anton Schwaighofer, Sebastian Mika, Antonius Ter Laak, Detlev Suelzle, Ursula Ganzer, Nikolaus Heinrich, and Klaus-Robert Müller. Estimating the domain of applicability for machine learning qsar models: a study on aqueous solubility of drug discovery molecules. *Journal of Computer-Aided Molecular Design*, 21(12):651–664, 2007. URL http://dx.doi.org/10.1007/s10822-007-9160-9http://download.springer.com/static/pdf/882/art%253A10.1007%252Fs10822-007-9160-9.pdf?auth66=1391651953_3b02d0898837e30891c01ext=.pdf.

Irene Luque Ruiz and Miguel Ángel Gómez-Nieto. Study of the applicability domain of the qsar classification models by means of the rivality and modelability indexes. *Molecules (Basel, Switzerland)*, 23 (11):2756, 2018. ISSN 1420-3049. 10.3390/molecules23112756. URL https://pubmed.ncbi.nlm.nih.gov/30356020https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278359/.

Kunal Roy, Supratik Kar, and Pravin Ambure. On a simple approach for determining applicability domain of qsar models. *Chemometrics and Intelligent Laboratory Systems*, 145:22–29, 2015. ISSN 0169-7439. https://doi.org/10.1016/j.chemolab.2015.04.013. URL http://www.sciencedirect.com/science/article/pii/S0169743915000969.

Pau Carrió, Marta Pinto, Gerhard Ecker, Ferran Sanz, and Manuel Pastor. Applicability domain analysis (adan): A robust method for assessing the reliability of drug property predictions. *Journal of Chemical Information and Modeling*, 54(5):1500–1511, 2014. ISSN 1549-9596. 10.1021/ci500172z. URL https://doi.org/10.1021/ci500172z.

Jochen Sieg, Florian Flachsenberg, and Matthias Rarey. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. *Journal of Chemical Information and Modeling*, 59(3):947–961, 2019. ISSN 1549-9596. 10.1021/acs.jcim.8b00712. URL https://doi.org/10.1021/acs.jcim.8b00712.

Markus C. Hemmer, Valentin Steinhauer, and Johann Gasteiger. Deriving the 3d structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy*, 19(1):151–164, 1999. ISSN 0924-2031. https://doi.org/10.1016/S0924-2031(99)00014-4. URL http://www.sciencedirect.com/science/article/pii/S0924203199000144.

Benjamin P. Brown, Jeffrey Mendenhall, and Jens Meiler. Bcl::molalign: Three-dimensional small molecule alignment for pharmacophore mapping. *Journal of Chemical Information and Modeling*, 59(2):689–701, 2019b. ISSN 1549-9596. 10.1021/acs.jcim.9b00020. URL https://doi.org/10.1021/acs.jcim.9b00020.

Jincai Yang, Cheng Shen, and Niu Huang. Predicting or pretending: Artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. *Frontiers in Pharmacology*, 11:69, 2020. ISSN 1663-9812. URL https://www.frontiersin.org/article/10.3389/fphar.2020.00069.

Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J. Dickson, Jose S. Duca, Viktor Hornak, David R. Koes, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLOS ONE*, 14(8):e0220113, 2019. 10.1371/journal.pone.0220113. URL https://doi.org/10.1371/journal.pone.0220113.

Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., 2016. ISBN 0128042915.

Z. Wang, H. Bian, S. G. Bartual, W. Du, J. Luo, H. Zhao, S. Zhang, C. Mo, Y. Zhou, Y. Xu, Z. Tu, X. Ren, X. Lu, R. A. Brekken, L. Yao, A. N. Bullock, J. Su, and K. Ding. Structure-based design of tetrahydroisoquinoline-7-carboxamides as selective discoidin domain receptor 1 (ddr1) inhibitors. *J Med Chem*, 2016. ISSN 1520-4804 (Electronic) 0022-2623 (Linking). 10.1021/acs.jmedchem.6b00140. URL http://www.ncbi.nlm.nih.gov/pubmed/27219676http://pubs.acs.org/doi/pdfplus/10.1021/acs.jmedchem.6b00140.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2019. ISSN 1549-960X. 10.1021/acs.jcim.8b00545. URL https://pubmed.ncbi.nlm.nih.gov/30481020.

M. J. O'Meara, A. Leaver-Fay, M. Tyka, A. Stein, K. Houlihan, F. DiMaio, P. Bradley, T. Kortemme, D. Baker, J. Snoeyink, and B. Kuhlman. A combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. *J Chem Theory Comput*, 11(2):609–622, 2015. ISSN 1549-9618 (Print) 1549-9618 (Linking). 10.1021/ct500864r. URL http://www.ncbi.nlm.nih.gov/pubmed/25866491http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4390092/pdf/nihms669655.pdf.

Shannon T. Smith and Jens Meiler. Assessing multiple score functions in rosetta for drug discovery. *PLOS ONE*, 15(10):e0240450, 2020. 10.1371/journal.pone.0240450. URL https://doi.org/10.1371/journal.pone.0240450.

Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, and R. Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–12, 2015. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). 10.1093/bioinformatics/btu626. URL http://www.ncbi.nlm.nih.gov/pubmed/25301850http://bioinformatics.oxfordjournals.org/content/31/3/405.full.pdf.

Faizan Sahigara, Kamel Mansouri, Davide Ballabio, Andrea Mauri, Viviana Consonni, and Roberto Todeschini. Comparison of different approaches to define the applicability domain of qsar models. *Molecules*, 17 (5), 2012. ISSN 1420-3049. 10.3390/molecules17054791.

P. Douglas Renfrew, Eun Jung Choi, Richard Bonneau, and Brian Kuhlman. Incorporation of noncanonical amino acids into rosetta and use in computational protein-peptide interface design. *PLOS ONE*, 7(3):e32637, 2012. 10.1371/journal.pone.0032637. URL https://doi.org/10.1371/journal.pone.0032637.

Jason W. Labonte, Jared Adolf-Bryfogle, William R. Schief, and Jeffrey J. Gray. Residue-centric modeling and design of saccharide and glycoconjugate structures. *Journal of computational chemistry*, 38(5): 276–287, 2017. ISSN 1096-987X 0192-8651. 10.1002/jcc.24679. URL https://pubmed.ncbi.nlm.nih.gov/27900782https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5182120/. 27900782[pmid] PMC5182120[pmcid].

Julia Koehler Leman, Rebecca F. Alford, and Jeffrey J. Gray. Rosetta-mpdock: A novel computational tool for protein-protein docking within the membrane bilayer. *Biophysical Journal*, 108(2):250a, 2015. ISSN 0006-3495. 10.1016/j.bpj.2014.11.1382. URL https://doi.org/10.1016/j.bpj.2014.11.1382. doi: 10.1016/j.bpj.2014.11.1382.

Barun Okram, Advait Nagle, Francisco J. Adrián, Christian Lee, Pingda Ren, Xia Wang, Taebo Sim, Yongping Xie, Xing Wang, Gang Xia, Glen Spraggon, Markus Warmuth, Yi Liu, and Nathanael S. Gray. A general strategy for creating "inactive-conformation" abl inhibitors. *Chemistry Biology*, 13(7):779–786, 2006. ISSN 1074-5521. https://doi.org/10.1016/j.chembiol.2006.05.015. URL https://www.sciencedirect.com/science/article/pii/S1074552106001840.

Yilin Meng, Cen Gao, David K. Clawson, Shane Atwell, Marijane Russell, Michal Vieth, and Benoît Roux. Predicting the conformational variability of abl tyrosine kinase using molecular dynamics simulations and markov state models. *Journal of Chemical Theory and Computation*, 14(5):2721–2732, 2018. ISSN 1549-9618. 10.1021/acs.jctc.7b01170. URL https://doi.org/10.1021/acs.jctc.7b01170. doi: 10.1021/acs.jctc.7b01170.

Scott A. Hollingsworth, Brendan Kelly, Celine Valant, Jordan Arthur Michaelis, Olivia Mastromihalis, Geoff Thompson, A. J. Venkatakrishnan, Samuel Hertig, Peter J. Scammells, Patrick M. Sexton, Christian C. Felder, Arthur Christopoulos, and Ron O. Dror. Cryptic pocket formation underlies allosteric modulator selectivity at muscarinic gpcrs. *Nature Communications*, 10(1):3289, 2019. ISSN 2041-1723. 10.1038/s41467-019-11062-7. URL https://doi.org/10.1038/s41467-019-11062-7.

A. C. Runcie, M. Zengerle, K. H. Chan, A. Testa, L. van Beurden, M. G. J. Baud, O. Epemolu, L. C. J. Ellis, K. D. Read, V. Coulthard, A. Brien, and A. Ciulli. Optimization of a "bump-and-hole" approach to

allele-selective bet bromodomain inhibition. *Chemical science*, 9(9):2452–2468, 2018. ISSN 2041-6520 2041-6539. 10.1039/c7sc02536j. URL https://pubmed.ncbi.nlm.nih.gov/29732121https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5909127/. 29732121[pmid] PMC5909127[pmcid] c7sc02536j[PII].

Daniel J. Urban and Bryan L. Roth. Dreadds (designer receptors exclusively activated by designer drugs): Chemogenetic tools with therapeutic utility. *Annual Review of Pharmacology and Toxicology*, 55(1):399–417, 2015. ISSN 0362-1642. 10.1146/annurev-pharmtox-010814-124803. URL https://doi.org/10.1146/annurev-pharmtox-010814-124803. doi: 10.1146/annurev-pharmtox-010814-124803.