

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Studying Proteins Implicated in Cancer with a Computational Toolbox

Permalink

<https://escholarship.org/uc/item/4rw973hd>

Author

Offutt, Tavina

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Studying Proteins Implicated in Cancer with a Computational Toolbox

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor
of Philosophy

in

Chemistry

by

Tavina Lynn Offutt

Committee in charge:

Professor Rommie Amaro, Chair
Professor Ruben Abagyan
Professor Elizabeth Komives
Professor Andrew McCammon
Professor Susan Taylor

2018

Copyright

Tavina Lynn Offutt, 2018

All rights reserved.

The Dissertation of Tavina Lynn Offutt is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2018

DEDICATION

This dissertation is dedicated to my one true love, my rock, and my husband, David Offutt Jr. Thank you for your unwavering support throughout my graduate school trajectory. I am grateful to you for always holding me accountable, and pushing me to achieve my dreams, and surpass my wildest expectations. You are my hero, and you always inspire me to be a better version of myself. I am fortunate to have had your support, guidance, and encouragement throughout this journey.

SEMPER FIDELIS

EPIGRAPH

When things go wrong as they sometimes will,
When the road you're trudging seems all uphill,
When the funds are low and the debts are high,
And you want to smile, but you have to sigh,
When care is pressing you down a bit,
Rest, if you must, but don't you quit.

Life is queer with its twists and turns,
As everyone of us sometimes learns,
And many a failure turns about,
When he might have won if he'd stuck it out.
Don't give up though the pace seems slow,
You might succeed with another blow.

Often the goal is nearer than,
It seems to a faint and faltering man,
Often the struggler has given up
When he might have captured the victor's cup;
And he learned too late when the night came down,
How close he was to the golden crown.

Success is failure turned inside out,
The silver tint of the clouds of doubt,
And you never can tell just how close you are,
It may be near when it seems so far,
So stick to the fight when you're hardest hit,
It's when things seem worst that you must not quit!

- Author Unknown

TABLE OF CONTENTS

SIGNATURE PAGE	iii
DEDICATION	iv
EPIGRAPH	iv
TABLE OF CONTENTS	vi
LIST OF FIGURES	viii
LIST OF TABLES	x
ACKNOWLEDGEMENTS	xi
VITA	xiv
ABSTRACT OF THE DISSERTATION	xv
CHAPTER 1: Introduction	1
CHAPTER 2: Molecular Dynamics of the R175H Mutant in the Full-Length p53 Tetramer Reveal Insight into the DNA Search and Recognition Mechanism	77
CHAPTER 3: Enhancing Virtual Screening Performance of Protein Kinases with Molecular Dynamics Simulations	118
CHAPTER 4: Knowledge-Based Methods to Train and Optimize Virtual Screening Ensembles	142
CHAPTER 5: Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands	160

LIST OF FIGURES

Figure 1.1: Six hallmarks of cancer	5
Figure 1.2: Schematic of ATP-mediated phosphorylation of protein kinases	16
Figure 1.3: Structure of the eukaryotic protein kinase domain.....	17
Figure 1.4: Assembly of the hydrophobic “spine” in active kinases	19
Figure 1.5: The eukaryotic protein kinase (ePK) protein kinome tree	21
Figure 1.6: Summary of the p53 pathway.....	28
Figure 1.7: Structure of full-length p53	30
Figure 1.8: Example free-energy landscape of proteins	39
Figure 1.9: Example force field equation used in molecular dynamics simulations	42
Figure 2.1: Full-length p53 System	82
Figure 2.2: Differences in the DNA binding mode between wildtype and R175H p53 ...	83
Figure 2.3: Effects of p53 R175H mutation on the L2 and L3 loops in the unique PC conformations	87
Figure 2.4: Solvent accessibility of p53 DBD	89
Figure 2.5: Footprint analysis of the CTD-DNA residues averaged across all three p53 systems.....	92
Figure 2.S1: Comparison between DNA binding modes of DBD.....	112
Figure 2.S2: Analysis of the Q180-R174 salt-bridge.....	113
Figure 2.S3: Root-mean-square fluctuations of DBD: compare wildtype to R175H with zinc.....	114
Figure 2.S4: Root-mean-square fluctuations of DBD: compare wildtype to R175H without zinc.....	115
Figure 2.S5: CTD footprint analysis based on DNA contacts	116
Figure 2.S6: DNA footprint analysis based on CTD contacts	117

Figure 3.1: Protein kinases involved in study.....	120
Figure 3.2: Comparison between how RMSD and POVME clusters the MD trajectory	123
Figure 3.3: ROC-EF of the cluster centroids and crystal structures against training set for each protein kinase.....	123
Figure 3.4: Trained ensemble sizes and crystal structures ROC-EF (fpf = 0.001) values against the training and test set across all six protein kinases	125
Figure 3.5: Docked pose of active compound, CHEMBL272309, reveals favorable interactions with RMSD centroid 3 (purple) and steric clashes (circled in left figure) with the crystal structure (pink)	128
Figure 3.S1: Comparison between the crystallographic and docked pose of the co-crystallized inhibitors.....	133
Figure 3.S2: The cluster centroids and crystal's AUC against the training set for each protein kinase	134
Figure 3.S3: Comparison between all structural selection methods.....	134
Figure 3.S4: The RMSD and POVME cluster centroids and randomly selected frames AUC values against the entire dataset of actives and decoys	135
Figure 3.S5: The RMSD and POVME cluster centroids and randomly selected frames ROC-EF values against the entire dataset of actives and decoys	135
Figure 3.S6: The trained ensemble sizes and crystal structures AUC values against the training and test set across all six protein kinases are shown	136
Figure 3.S7: The cluster centroids and crystal's AUC against the test set for each protein kinase	136
Figure 3.S8: The cluster centroids and crystal's ROC-EF against the test set for each protein kinase	137
Figure 3.S9: ROC-EF values at a later false positive fraction (fpf = 0.01) against the training and test set for each ensemble combination using RMSD centroids across all six protein kinases	138
Figure 3.S10: ROC-EF values at a later false positive fraction (fpf = 0.01) against the training and test set for each ensemble combination using POVME centroids across all six protein kinases.....	138
Figure 3.S11: Comparison between crystal structures and MD conformations	139

Figure 4.1: AUC and EF histograms.....	148
Figure 4.2: Training set performance as a function of ensemble size for three proteins using DUD-E.	149
Figure 4.3: Receiver operating characteristic (ROC) curves for ensembles trained to maximize the AUC of the ROC curve	150
Figure 4.4: Receiver operating characteristic (ROC) curves for ensembles trained to the EF at a FPF of 0.001	151
Figure 4.5: Percentages of compounds whose graph frameworks (FWs) are unique to, and shared between, training and test sets	152
Figure 4.6: Training method schematic: selecting the best performing ensemble from three target conformers	152
Figure 5.1: Computational/experimental protocol used to identify novel estrogen-receptor ligands	162
Figure 5.2: ROC curves associated with each of the three high-performing virtual screens	163
Figure 5.3: Crystallographic pose of estradiol	166
Figure 5.4: Binding poses	167

LIST OF TABLES

Table 2.1: Hydrogen Bonding Interactions between loops in Monomers B and C	85
Table 2.S1: Summary of Simulated Model Systems	109
Table 2.S2: Salt Bridge Footprint Analysis of CTD-DNA.....	109
Table 2.S3: L1/S3 Pocket Open Ratios in MD Simulations.....	112
Table 3.1: Protein Kinase Systems Setup for MD Simulations and VS Training	121
Table 3.2: Global Performance of the Virtual Screen (AUC) of the Optimal Trained Ensemble against the Test Set.....	125
Table 3.3: Early Chemical Enrichment of Actives (ROC-EF _{f_{pf}=0.001}) of the Optimal Trained Ensemble against the Test Set	125
Table 3.S1: Number of Crystal Structures used for PCA Comparison between MD and Crystal Structure Conformations	140
Table 4.1: Summary of Structures Used To Construct Ensembles	147
Table 4.2: AUC Values Determined on Training and Test Sets of Best Performing Ensembles Selected To Maximize AUC	150
Table 4.3: EF at FPF of 0.001 Determined on Training and Test Sets of Best Performing Ensembles Selected To Maximize EF at FPF of 0.001	151
Table 5.1: High-Affinity Compounds Found by Docking into ER α Structures in Both the Antagonist- and Agonist-Bound Conformations, Sorted by the Experimentally Measured ER α K_i	164
Table 5.2: Chemical-Diversity Analysis Using Molecular Graphs	166
Table 5.S1: Additional compounds docked and evaluated using HTVS-SP-XP-NN2 into an ER α structure in the antagonist-bound conformation	171
Table 5.S2: Additional compounds docked and evaluated using HTVS into an ER α structure in the agonist-bound conformation	172
Table 5.S3: Additional compounds docked and evaluated using HTVS-SP-XP-NN1 into an ER α structure in the agonist-bound conformation	173
Table 5.S4: Relative binding affinity (RBA) values.....	174
Table 5.S5: Chemical diversity analysis using cumulative frequency scaffold plots.....	175

ACKNOWLEDGEMENTS

I would like to acknowledge my research advisor, Professor Rommie Amaro, for her guidance, support, and mentorship throughout my graduate research. I am appreciative of your willingness to allow me flexibility in my research, and your optimism and enthusiasm when providing direction on my research projects. Thank you for being an excellent role model; a person I can look up to as I pursue a career in scientific research. There is so much that I have learned from you that I plan to carry into my own research career.

I would like to thank all of my committee members, Professor Elizabeth Komives, Professor Susan Taylor, Professor Andrew McCammon, Professor Roy Wollman, Professor Arnold Rheingold, and Professor Ruben Abagyan. Thank you all for your advice and candor on my research projects. I especially would like to thank Professor Komives for being a great mentor on how to navigate graduate school. You took me under your wing and supported me, and for this I am forever grateful.

I would like to thank Amaro group members, Dr. Robert Swift and Dr. Özlem Demir for being awesome mentors on my research projects. Robert, thanks for all the discussions we had pertaining to the protein kinases work. You are such a great teacher, and I learned a great deal from you. Ozlem, thanks for your guidance on the p53 work, and for your willingness to share your knowledge. Both of you are brilliant scientists, and I am very fortunate to have worked closely with both of you.

I would also like to thank all of my lab mates in the Amaro group, both past and present. Thank you for being such supportive and friendly lab mates, and for all the

scientific discussions. I am grateful to have worked with such great people, and appreciate all that I have learned from each and every one of you. A special thanks to Victoria Feher for your willingness to share your expertise on protein kinases, and Robert Malmstrom for your expertise and advice on Markov State Models. I would also like to thank Dr. Jamie Schiffer who has been such a great friend. I feel like we tag teamed graduate school together, and I couldn't have done it without you. Also, thanks so much for your support during this thesis writing process.

I would like to thank Professor Seth Cohen for his guidance, support, and mentorship as a co-advisor during my initial graduate research projects. I would also like to thank the Cohen group members for their support and friendship.

I would like to acknowledge family and friends who have provided emotional and mental support throughout this process. First, I would like to thank my husband for his support and personal investment in my accomplishing this monumental goal. I could not have done it without you. I would like to thank my parents for instilling academic excellence in me at a young age, and my siblings for setting a great example for me and their love and support.

Also, I would like to acknowledge great mentors I have had throughout my entire life. I would like to thank Cassaundra Taylor, my fifth grade teacher, for being such a great role model early in my life, and for always encouraging me to dream big. I would like to thank Dr. Lisa Rhodes and Kimberly Scott, for their love, support, and personal investment in my personal growth. I am forever in debt to you all, and I hope that I am as great of a mentor to future young girls and women that you are to me.

Chapter 2, in part is currently being prepared for submission for publication of the material. Offutt, Tavina L.; Jeong, Pek U.; Demir, Özlem; Amaro, Rommie E. The dissertation author is the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Enhancing Virtual Screening Performance of Protein Kinases with Molecular Dynamics Simulations 2016. Offutt, Tavina L.; Swift, Robert, V.; Amaro, Rommie E., J Chem Inf Mod, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Knowledge-Based Methods To Train and Optimize Virtual Screening Ensembles 2015. Swift, Robert V.; Jusoh, Siti A.; Offutt, Tavina L.; Li, Eric S.; Amaro, Rommie E., J Chem Inf Mod, 2016. The dissertation author was a secondary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands 2015. Durrant, Jacob D.; Carlson, Kathryn E.; Martin, Teresa A.; Offutt, Tavina L.; Mayne, Christopher G.; Katzenellenbogen, John A.; Amaro, Rommie E., J Chem Inf Mod, 2015. The dissertation author was a fourth investigator and author of this paper.

VITA

Bachelor of Science, Spelman College	2010
Massachusetts Institute for Technology B-cubed Post baccalaureate	2012
Research Assistant, University of California, San Diego	2012-2017
Master of Science in Chemistry, University of California, San Diego	2014
Doctor of Philosophy, University of California, San Diego	2018

ABSTRACT OF THE DISSERTATION

Studying Proteins Implicated in Cancer with a Computational Toolbox

by

Tavina Lynn Offutt

Doctor of Philosophy in Chemistry

University of California, San Diego, 2018

Professor Rommie Amaro, Chair

Cancer formation is a complex, multi-step process that allows cells to grow abnormally and potentially invade and spread throughout the body. A single genetic or structural alteration of a single protein in a cellular physiological process is enough to stimulate cancer formation. In treating cancer, a ‘targeted therapy’ approach is becoming

increasingly common, where we can develop drugs that specifically target these altered proteins implicated in cancer. Two proteins that are heavily involved in several human cancers are protein kinases and p53, which are the focus of this dissertation work. I chose to use molecular dynamics simulations and *in silico* virtual screening, two methods from the computational chemistry toolbox in studying protein kinases and p53. I demonstrate that performing molecular dynamics is worthwhile in conducting virtual screens against protein kinases, because it may result in that at least one conformation is more predictive than the crystal structure. I also reveal key insight into the transcriptional activation mechanism of p53, and show how this mechanism is altered as a result of the R175H cancer mutation.

Chapter 1: Introduction

Synopsis

In this dissertation, proteins that are implicated in human cancer are explored using computational techniques, specifically molecular dynamics and virtual screening. As an introduction to this dissertation, background on cancer is introduced. Following an introduction to cancer, two cancer drug targets, protein kinases and tumor suppressor p53, are introduced. Lastly, molecular dynamics is discussed in more detail and its utility in studying protein kinases and p53 are highlighted.

For the background on cancer (*“Introduction to Cancer”*), the epidemic of cancer is briefly discussed to highlight the broad impact of this dissertation work in *Section i*. In *Section ii*, the biology involved in cancer formation is detailed to introduce the reader to the many cellular processes impacted in tumor formation and progression (tumorigenesis). In *Section iii*, the current cancer treatments are reviewed. This section also discusses the limitations with traditional cancer treatments, and how ‘targeted cancer therapy’ overcomes these limitations.

Next, protein kinases are introduced (*“Protein Kinases”*). In *Section i*, the role protein kinases plays in cellular processes are introduced. In *Section ii*, the structure of the protein kinase domain (the domain that contains the active site) is detailed. Next, the reader is introduced to the classification of protein kinases in the human genome (*Section iii*). In *Section iv*, protein kinases implication in cancer is discussed, followed by a brief introduction of each of the six protein kinases studied in this dissertation (*Section v*).

Following discussion of protein kinases, the tumor suppressor p53 is introduced (*“Tumor Suppressor, p53”*). In *Section i*, the role p53 plays in biology is discussed. In

Section ii, the structure of full-length p53 and the function of each domain are detailed. Next, the implication of p53 in cancer is presented (*Section iii*), followed by a discussion of the various therapeutic approaches against p53 (*Section iv*).

Lastly, molecular dynamics simulations use in exploring protein dynamics is introduced (*“Molecular Dynamics Applied to Biological Macromolecules”*). First, the reader is introduced to the dynamic behavior of proteins to provide context for the reason we use molecular dynamics in studying protein dynamics *Section i*. Next *in Section ii*, the theory in molecular dynamics is discussed, followed by molecular dynamics utility in drug discovery efforts, specifically in virtual screening methods (*Section iii*).

Introduction to Cancer

(i) Cancer Epidemics

Cancer diseases are the leading causes of morbidity and mortality worldwide. According to the National Cancer Institute (NCI), there was an estimate of 1,685,210 new cases of cancer in the United States, and 595,690 people were presumed to die from the disease in 2016. In 2014, approximately 14.5 million people were diagnosed with cancer, and this number is expected to rise to almost 19 million by 2024.² Due to this global burden, there are great research efforts to both treat and prevent cancer diseases. While there have been significant advancements in cancer research leading to a decline in the number of cancer deaths each year, there are still several areas of improvement in developing cancer treatments, which will be discussed in more detail in *Section iii*.

(ii) Cancer Biology

Cancer is a group of diseases that result from transformed normal cells that grow and multiply uncontrollably. This transformation of normal cells to cancer cells is a multistep process that involves genetic alterations. Six hallmarks or essential alterations of cell physiology must occur in cancer formation: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Figure 1.1).¹ Cancer studies suggest that these six capabilities are shared in common by all human types of cancer.

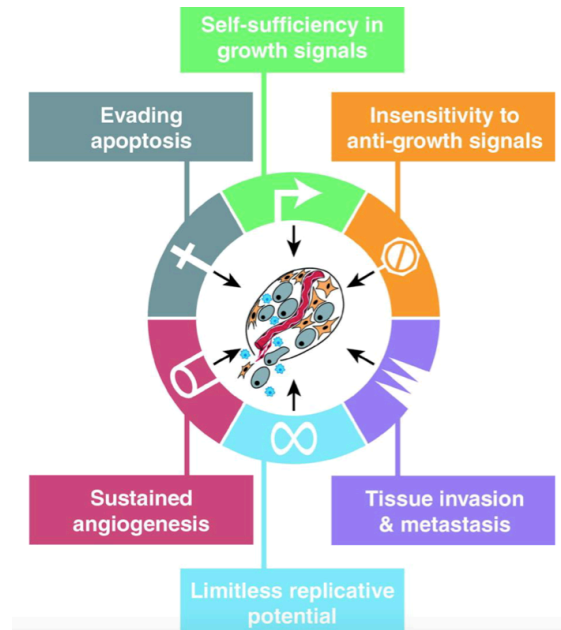


Figure 1.1: Six hallmarks of cancer. Most cancers must acquire the same set of functional capabilities during tumor formation. Adapted from Hanahan and Weinberg.¹

Acquired growth signaling autonomy was the first hallmark of cancer identified by researchers. In normal cell growth and division, transmission of mitogenic growth signals (GS) are required to allow cells to move from a quiescent to an active proliferative state. This is accomplished in three steps: (i) one type of cell makes soluble proteins, mitogenic growth factors (GF), (ii) these GFs bind to another cell surface receptor (termed a growth factor receptor), and (iii) this GF receptor-binding event leads to intracellular stimulatory signals that induces cell proliferation. Cancer cells achieve GS autonomy via alteration of any of these three steps in mitogenic growth signaling.

Three common molecular strategies for achieving GS autonomy involve alteration of extracellular GS, GF receptors, or intracellular signaling cascades that transduce the GS signals into action. Cancer cells can manufacture their own GFs, preventing

dependence on GFs from other cells. Also, the GF receptors, which are often tyrosine kinases (*further discussion of protein kinases activity in cancer will be discussed in Protein Kinases Section iv*), are typically overexpressed in cancers, allowing ambient levels of GF that normally would not trigger proliferation to bind cancer cell GF receptors, triggering proliferation. Also, cancer cells may favor expression of extracellular matrix receptors (integrins) that transmit progrowth signals. Lastly, proteins that are involved in downstream cytoplasmic circuitry induced by GF receptor binding are altered in cancer cells. For example, Ras proteins are structurally altered in 25% of human cancers, and this structural deviation allows Ras to release mitogenic signals without GF receptor binding.³

The second hallmark of cancer is evasion of anti-proliferative signals, which maintains cellular quiescence and tissue homeostasis in normal tissue. Much of the antigrowth signaling is associated with the cell cycle clock. Cells monitor their external environment during the G1 phase (the first growth period of the cell cycle, where the cell grows and cytoplasmic organelles are replicated) of its growth cycle, where sensed signals dictate whether to proliferate, to be quiescent, or to enter into a postmitotic state. Majority of the anti-proliferative signals are channeled through the retinoblastoma protein (pRb), and its relatives, p107 and p130. Hypophosphorylation (removal of phosphate groups) of pRb leads to sequestration with E2F transcription factors, which blocks E2F activation of genes that promote cell progression from G1 to S phase, thereby inhibiting proliferation.⁴ pRb is inactivated via phosphorylation by cyclin:cyclin dependent kinase (CDK) complexes, which leads to cell progression through the G1 phase. The soluble signaling protein, TGF β , disrupts phosphorylation of pRb by inducing synthesis of

p15^{INK4B} and p21 proteins, which block the cyclin:CDK complexes responsible for pRb phosphorylation.^{5,6} In this way, TGFβ governs the pRb signaling circuit.

In human tumors, the pRb signaling circuit can be disrupted in a variety of ways,⁷ allowing cancer cells to avoid antigrowth signals. TGFβ receptors may be downregulated or mutated, causing cancer cells to lose TGFβ responsiveness. The Smad4 cytoplasmic protein, which transduces signals from TGFβ-TGFβ receptor binding, may be mutated, thereby disrupting the TGFβ-mediated pRb signaling circuit.⁸ The locus encoding p15^{INK4B} may be deleted⁹ or mutated,¹⁰ thereby allowing cyclin:CDK complex formation. Lastly, the function of pRb may be disrupted via mutation or sequestration by viral oncoproteins, such as the E7 oncoprotein of human papillomavirus.¹¹ Through multiple avenues, the pRb antigrowth signaling circuit is disrupted in a majority of human cancers.

There exists a significant amount of evidence that acquired resistance toward apoptosis is a third hallmark of all cancer types. There are a variety of physiologic signals that triggers apoptosis, which unfolds in a precise series of steps, resulting in the cell being engulfed by nearby cells within 24 hours.¹² The apoptotic machinery consists of sensors whose job is to monitor the extracellular and intracellular environment for conditions that influence whether a cell should live or die. These sensors signal effector molecules, which bind either death or survival factors. There are various pathways in which apoptosis can be initiated, in which we will only discuss two. One pathway involves the insulin growth factor receptor 1 (IGF-1R). When the sensor molecules sense normal extracellular and intracellular conditions, the survival factors, IGF1 and IGF2, are signaled to bind IGF-1R, resulting in downstream antiapoptotic survival signaling.^{13,14}

Another pathway involves the p53 tumor suppressor protein, which upregulates the expression of Bax (a member of the Bcl-2 family of proteins that has proapoptotic functions), which then stimulates the mitochondria to release cytochrome C, a potent catalyst of apoptosis.¹⁵

Cancer cells can acquire resistance to apoptosis through different strategies. Disruption of either pathway briefly described above can disrupt the apoptotic machinery. Mutation of the IGF-1R can prevent initiation of the antiapoptotic survival signaling cascade. Also any structural alteration of proteins involved in the downstream antiapoptotic signaling can disrupt this pathway. For example, the PI3 kinase-AKT pathway, which transmits antiapoptotic survival signals, is involved in abrogating apoptosis in several human tumors. The most common way cancer cells develop resistance to apoptosis is through mutation of the p53 tumor suppressor gene, thereby preventing apoptosis even when the cell is damaged. Altering the apoptotic machinery is essential for tumor progression, which can be achieved through various strategies.

In addition to disruption of cell-to-cell signaling as described in the first three hallmarks, tumor cells must acquire immortality, which brings us to the fourth hallmark of cancer. Work performed by Hayflick demonstrated that normal cells have a finite replicative potential.¹⁶ Once normal cells progress through a predetermined number of doublings (approximately 60-70), they stop growing – a process termed senescence. Senescence can be circumvented through p53 or pRb tumor suppressor proteins, allowing the cells to continue multiplying until they enter a second state termed crisis.

Massive cell death, end-to-end fusion of chromosomes resulting in karyotypic disarray, and immortality of 1 in 10^7 cells are all characterizations of the crisis state.¹⁷

Telomeres, the ends of chromosomes, are used as a counting device for cell generations. With each cell replication cycle, telomeres become shortened. This erosion of telomeres disrupts their ability to protect the ends of chromosomal DNA, resulting in the end-to-end chromosomal fusions yielding the karyotypic disarray associated with the crisis state, thereby leading to massive cell death.¹⁸

Cancer cells avoid the crisis state through telomere maintenance, which is evident in almost all types of cancers.¹⁹ Most cancer cells (85%-90%) obtain this through upregulated expression of the telomerase enzyme, which adds hexanucleotide repeats onto the ends of telomeric DNA.²⁰ Other cancer cells achieve telomere maintenance by an invented activation mechanism that involves recombination-based interchromosomal exchanges of sequence information.²¹ Through either mechanism, the maintenance of telomeres in cancer cells permits unlimited multiplication of descendent cells.

Extensive and compelling experimental studies suggest that induction of angiogenesis, the fifth hallmark of cancer, may be an early to midstage event in many cancers. Angiogenesis, the growth of new blood vessels, is essential for cell survival and function as it provides oxygen and nutrients to cells. This process is carefully regulated by positive and negative signals that either encourage or block angiogenesis. There is over a dozen of growth factors that bind transmembrane tyrosine kinase receptors displayed on the surface of endothelial cells, inducing angiogenesis-initiating signals.^{22, 23} Alternatively, there are endogenous inhibitor proteins that block angiogenesis.

In addition to soluble factors and their receptors, integrin signaling is also involved in angiogenesis regulation. Interference with signaling from sprouting capillaries, one class of integrins, can inhibit angiogenesis.^{24, 25} Extracellular proteases,

proteolytic enzymes, together with proangiogenic integrins, help dictate the invasive capability of angiogenic endothelial cells.²⁶

Since angiogenesis is critical for cell survival and function, cancer cells have developed a strategy to induce angiogenesis during tumor progression. Tumors can activate angiogenesis by shifting the balance between angiogenic inducers and inhibitors (Hanahan and Folkman, 1996). Increased gene expression of the transmembrane tyrosine kinase receptors that bind soluble growth factors is one common strategy for this shift. Another strategy involves downregulation of endogenous angiogenic inhibitors. The mechanisms involved in shifting the balance between angiogenic inducers and inhibitors in cancer cells are still not completely elucidated. However, studies have shown how gene alteration may occur. For example, Dameron et al. found that p53 regulates the angiogenic inhibitor, thrombospondin-1.²⁷ This loss of p53 function, which is common in human tumors, would then lead to downregulation of thrombospondin-1, allowing angiogenesis to take place. There are other examples, which all suggest that cancer cells use different molecular strategies to activate the angiogenic switch.

The final hallmark of cancer, which depends upon the other five hallmark capabilities, involves invasion and metastasis. In these two processes, tumor cells produce cells that move and invade adjacent tissues, thereby succeeding in finding new colonies. These distant settlements are the cause of about 90% of human cancer deaths.²⁸ Invasion and metastasis are very complex processes that involve several different proteins such as proteases, cadherins, cell-cell adhesion molecules (CAMs), and integrins. While research studies have shown that these protein expression levels are

altered during tumor invasion and metastasis, the mechanistic role of these proteins remain incompletely understood.^{24, 26, 29-39}

With the large number of cancer diagnosis, there are significant efforts to improve drug discovery against these diseases. Since tumor progression is a complicated multi-step process, this poses many challenges in treating and preventing cancer. However, due to the hallmarks involved in cancer, there are several avenues available in targeting cancer with drug molecules. While the current commonly used treatment aims to kill cancer cells, even though not specifically, more personalized treatment is becoming more attractive. Due to our increasing knowledge of how tumors progress at a molecular level, we are now able to identify particular proteins that are promoting cancer, and specifically target them with drugs.

(iii) Treatments for Cancer

The landscape of cancer therapies has changed dramatically over the past five decades. Initially, more classical treatments such as surgery, radiation, chemotherapy, and endocrine therapy, were used to halt tumor growth. However, as researchers have elucidated the various molecular features involved in tumor growth (as discussed previously in *Introduction to Cancer Section ii*), a more targeted therapy approach has emerged. While the use of these targeted therapies alone and in combination with the classical approaches has increased the effectiveness of cancer treatments, due to various limitations, a need for improved treatments still exist.

Surgery is the primary form of treatment when the cancer is a solid localized tumor. During surgical procedures, the solid tumor is removed along with surrounding

normal cells and in some cases, surrounding lymph nodes.⁴⁰ This treatment is extremely effective in that 100% of the removed cancerous cells are killed.⁴¹ However, this does not guarantee removal of all cancer cells. Also, surgery may not be effective for certain types of cancers such as metastatic cancers and leukemia, where the cancer is found throughout the blood and is not a solid localized tumor. Therefore, it is common practice to perform surgery in combination with radiation and chemotherapy.

Radiation involves the use of ionizing radiation at high doses to kill cancer cells and shrink tumors. This form of treatment may be used before, during, or after surgery. Chemotherapy treatment encompasses two classes of drugs: alkylating agents and antimetabolites, which disrupt biological processes essential for cell division in cancer cells.⁴² Specificity and selectivity are major limitations of radiation therapy and chemotherapy. Both forms of treatment are unable to select for cancer cells only, thereby affecting normal cells, leading to unwanted side effects.

Endocrine or hormone-based therapies involve the manipulation of the endocrine system by administering either exogenous hormones or drugs that inhibit the production or activity of hormones implicated in cancers.⁴³ Endocrine therapy consists of various medication strategies, and was first applied to breast cancer patients, where ~80% of breast cancers are hormone-dependent. Selective estrogen receptor modulators (SERMs) block hormones from attaching to cancer cells; the most commonly used SERM is tamoxifen which competitively inhibits coactivators from binding the estrogen receptor. Aromatase inhibitors (drugs that disrupt aromatase enzyme function, thereby inhibiting estrogen production) have become the state-of-the-art treatment for estrogen-dependent breast cancer due to their favorable toxicity profile, unlike tamoxifen.⁴⁴ While hormone-

based therapies have proven more efficacious in comparison to chemotherapy for example, they are only able to treat cancers that are hormone-dependent, limiting their use.

Due to the elucidation of the molecular characteristics of cancer cells, new therapeutic strategies that target these specific molecular features have been developed. This has led to the era of ‘targeted therapy’, where medicines block cancer cell proliferation by interfering with molecules (growth factors, signaling molecules, cell-cycle proteins, apoptosis modulators, and molecules that promote angiogenesis) needed for carcinogenesis and tumor growth, as opposed to disrupting all rapidly dividing cells as seen in chemotherapy. The most successful examples of targeted therapies are chemical entities that target a protein or enzyme that carries some genetic or structural alteration in cancer cells and not normal cells. These chemical entities may be in the form of antibodies, small molecules, antiangiogenics, or viral vectors. However, we limit our discussion to the main categories of targeted therapies, which include monoclonal antibodies and small molecules.

The first demonstration of successful targeted therapy involved the HER-2/neu protein, a protein that belongs to a family of four transmembrane receptor tyrosine kinases that mediate cell growth, differentiation, and survival.⁴⁵ HER-2/neu protein is overexpressed in 20%-25% of breast cancers.⁴⁶ The monoclonal antibody, trastuzumab, received regulatory approval in treating Her-2 + positive breast cancer patients either in isolation or combination with chemotherapy.

One of the most successful molecular targeted therapeutic is imatinib mesylate (Gleevec), which is an inhibitor of the kinase BCR-Abl, a protein that promotes

tumorigenesis in chronic myeloid leukemia.⁴⁷ It also inhibits the KIT tyrosine kinase and platelet derived growth factor receptor- β in the treatment of gastrointestinal stromal tumours and hypereosinophilic syndrome.

While targeted therapies allow us to treat cancer by targeting specific molecules that are altered in cancer, there still remain limitations with these targeted approaches. For starters, obtaining drug selectivity may be challenging for certain protein classes that have similar active sites. For example, different protein kinases have high sequence conservation and similar architecture, thus achieving drug selectivity is a major challenge in the design of protein kinase inhibitors. Therefore, the drug may bind additional off-target proteins, leading to adverse side effects. Drug resistance is another challenge where the targeted protein may develop novel mutations. For example, p53 cancer mutations are the most common genetic event in human cancer.⁴⁸⁻⁵¹ These new mutations will shift the structure of the drug target, thereby preventing the drug from binding. The use of computational tools and models can overcome these limitations by allowing us to understand the dynamic properties of drug targets at an atomic level, and using this information in designing drug molecules.

Computational methods such as molecular dynamics (MD) simulations can use a 3-dimensional structure of the drug target and apply physics-based principles to simulate the target through time. We can use this dynamics information to enhance drug selectivity. While the structural architecture of protein kinases for example may be similar, their dynamic characteristics should differ and may even reveal novel pockets to target with drugs that are specific to one particular protein kinase. Computational approaches can also model mutant forms of proteins that yield drug resistance. MD

simulations can be performed on both the normal wildtype and mutant forms of a drug target such as p53. These MD snapshots can reveal unique mutant conformations, which can be used in drug discovery methods targeting these drug resistant mutants. Later in *Section (iii) of Molecular Dynamics Applied to Biological Macromolecules*, there will be more detailed discussion on the use of MD simulations in modeling protein dynamics and its application to drug discovery.

This dissertation work focuses on two important classes of enzymes that are heavily involved in tumorigenesis: protein kinases and p53. Protein kinases phosphorylate substrates in several cellular processes, many of which are processes that are impacted in tumorigenesis. Therefore, protein kinases are the most sought-after targets for cancer treatments.⁵² The ‘guardian of the genome’ as it is commonly referred to as, p53, functions as a tumor suppressor. It is the most frequently mutated gene in human cancer, highlighting its potential as a cancer drug target. Both of these cancer targets will be discussed in detail next, highlighting their biological function, structure, and implication in cancer. Proceeding discussion of protein kinases and p53, the role of computational modeling in understanding their dynamic behavior and its application to drug discovery efforts will be presented.

Protein Kinases

(i) Biology of Protein Kinases

Protein kinases are a large and diverse class of proteins that mediate most of the signal transduction in eukaryotic cells via phosphorylation of substrates (Figure 1.2).⁵³ Adenosine triphosphate (ATP)-mediated phosphorylation occurs on tyrosine, threonine, and serine residues, with tyrosine being the dominant phosphorylation site.⁵⁴ Through modification of substrate activity, protein kinases control many cellular processes including: metabolism, transcription, cell cycle progression, cytoskeletal rearrangement and cell movement, apoptosis, and differentiation.⁵⁵ Protein phosphorylation also plays a role in intercellular communication during development, physiological responses and in homeostasis, and in the nervous and immune systems. There exists a large amount of studies on protein kinases, as they are among the largest families of genes.⁵⁶⁻⁶⁰

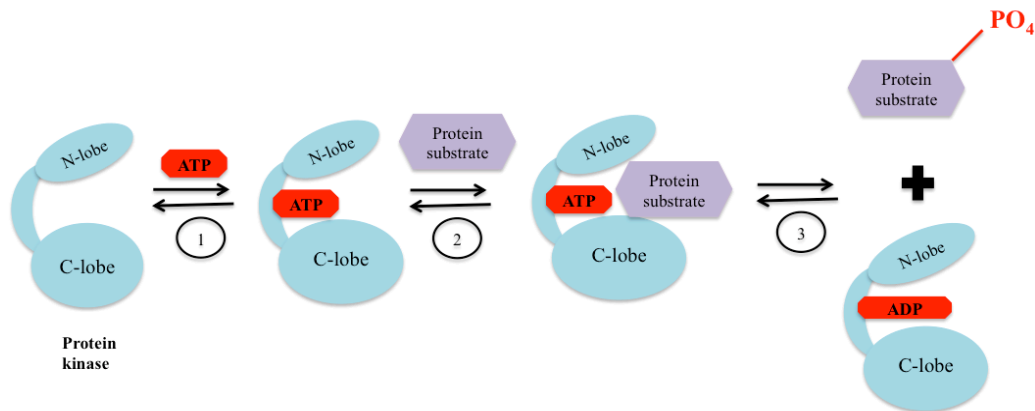


Figure 1.2: Schematic of ATP-mediated phosphorylation of protein kinases. First, an ATP molecule binds the active site of the protein kinase (1). The protein substrate binds near the ATP-binding site, where the γ phosphate group from ATP is transferred to either a threonine, tyrosine, or serine residue on the protein substrate (2), resulting in a phosphorylated protein substrate and adenine disphosphate (ADP)-bound protein kinase (3).

(ii) Structure of Protein Kinase Domain

The canonical protein kinase domain contains approximately 250 amino acids, is highly flexible, and consists of two lobes.⁶¹ The lobe located at the N-terminus (N-lobe) and the other lobe at the C-terminus (C-lobe) is separated by a cleft that contains the catalytic site, where ATP binding and phosphate transfer takes place (Figure 1.3). There is a short linker known as the hinge region between the two lobes, which forms hydrogen-bonding interactions with the adenine ring of ATP. This overall spatial architecture and general function of protein kinases is highly conserved across all families of protein kinases.

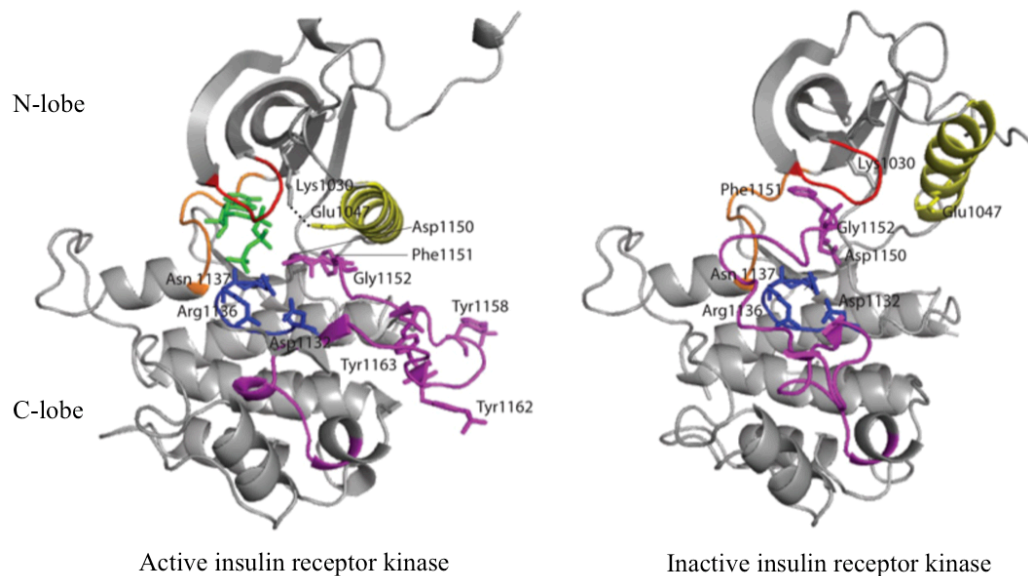


Figure 1.3: Structure of the eukaryotic protein kinase domain. The conformation of the active (PDB 1IR3) and inactive (PDB 1IRK) states of the insulin receptor kinase are shown. The kinase domain is depicted as a silver ribbon, with the C-helix, P-loop, hinge region, catalytic loop, and activation loop highlighted in yellow, red, orange, blue, and purple respectively. The binding of an ATP analog, adenylylimidodiphosphate (AMP-PNP), is shown in green sticks in the active conformation. The DFG motif is shown as purple sticks, where its orientation is away from the ATP-binding pocket in the active state and flipped into the ATP-binding pocket in the inactive state. The dashed lines indicate the salt bridge between Glu1047 in the C-helix and Lys1030 in the β 3-strand. Modified from Duong-Ly and Peterson.⁵⁵

The smaller N-lobe is made up of mostly beta strands (β 1- β 5), with one long alpha helix, known as the C-helix (Figure 1.3). Experimental site-directed mutagenesis studies suggest that protein-protein interactions take place in this region.⁶² β strands 1 and 2 contain the conserved glycine-rich sequence, which is commonly termed the phosphate-binding loop (P-loop). The P-loop coordinates one of the ATP phosphates, and is thought to contribute in several ways to protein kinase function.⁶³ There is a salt-bridge between a glutamic acid residue (Glu) at the beginning of the C-helix and a lysine residue (Lys) in the β 3-strand that is conserved in the active conformation, and missing in some inactive conformations of protein kinase crystal structures. Protein kinases are most sensitive to mutations of this Lys residue, which contains the α - and β -phosphoryl groups of the bound ATP.^{64, 65}

The larger C-lobe is predominately α helical, and contains residues that form interactions with the phosphate acceptor (Figure 1.3). There are several residues that interact with the triphosphate group of ATP in this lobe. The catalytic loop (residues 165 to 171 are directly involved in catalysis) is located between β strands 6 and 7, and forms the base of the active site. The C-lobe also contains the activation loop, which is flanked by the conserved DFG (Asp184-Phe185-Gly186) moiety, which can adopt a ‘DFG-in’ or ‘DFG-out’ conformation (Figure 1.3). Protein kinase inhibitors are classified based on the conformation the DFG moiety adopts when the inhibitors are bound (*further discussion of inhibitor types will take place in Section iv*).

The Asp184 in the DFG motif is one of the most important residues for catalysis. It is a coordination ligand of a magnesium ion, which positions the phosphate groups of

ATP for catalysis. The DFG phenylalanine forms hydrophobic contacts with the C-helix and the His-Arg-Asp (HRD) motif (conserved motif throughout the protein kinases family) in the catalytic loop. The DFG phenylalanine is also a member of the hydrophobic spine (Leu106-Leu95-Phe185-Tyr164), which is ordered in the active conformation and disordered in the inactive conformation of protein kinases (Figure 1.4).⁶⁶ The DFG glycine is highly conserved across protein kinases, and Kornev et al has shown that this residue acts as a bi-positional switch that reorients the DFG aspartate into active and inactive positions.⁶⁶

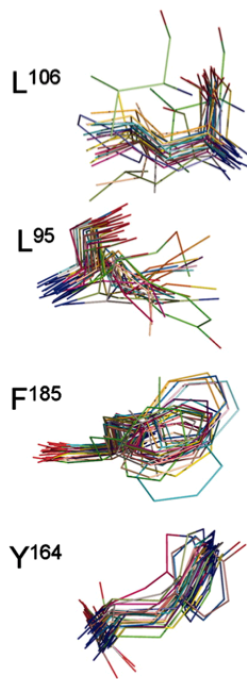


Figure 1.4: Assembly of the hydrophobic “spine” in active kinases. Alignment of the hydrophobic spine, which consists of four residues corresponding four PKA residues, L⁹⁵, L¹⁰⁶, Y¹⁶⁴, and F¹⁸⁵, are shown. Adapted from Kornev et al.⁶⁰

(iii) Human Kinome

A comprehensive genomewide study found that there are ~518 protein kinases in humans,⁶⁷ comprising ~1.7% of the human genome, and ~20 lipid kinases^{68, 69} that make up the human kinome. Of the protein kinases, 478 contain a eukaryotic kinase (ePK) domain and the 40 remaining kinases are classified as atypical protein kinases (aPKs) because they lack sequence similarity to the ePK domain, but they have kinase activity.⁶⁷ The ePKs are classified into eight major groups (Figure 1.5): (i) tyrosine kinases (TK), (ii) protein kinases A, G, and C (AGC), (iii) cyclin-dependent kinases (CDKs) and CDK-like kinases (CMGC), (iv) serine/threonine kinases (TKL), (v) kinases homologous to yeast proteins STE20, STE11, and STE7 (STE), (vi) casein kinase 1 and homologous kinases (CK1), (vii) kinases involved in calcium signaling (CAMK), and (viii) receptor guanylyl cyclases (RGC), in which this dissertation focuses on protein kinases within the first three kinome groups.

The TK group consists of receptor tyrosine kinases (RTKs) and cytosolic tyrosine kinases. RTKs are transmembrane proteins, with an extracellular domain that binds ligands that transmit signals across the cell membrane into the cytoplasm. Examples of RTKs include the insulin receptor (IR) and the closely related insulin-like growth factor 1 receptor (IGF1R), the human epidermal growth factor receptor (HER/EGFR), the platelet-derived growth factor receptors (PDGFRs), and the fibroblast growth factor receptors (FGFRs). All of these RTKs have been associated with cancer progression, where alteration of these RTKs allows cancer cells to acquire growth-signaling autonomy (*recall the first hallmark of cancer identified in Introduction to Cancer Section ii*). The remainder of the TK is made up of soluble protein kinases in the cytoplasm. Examples

include Src (the first identified protein kinase), Abl, and JAK kinases. It is important to note that the TK group, specifically RTKs, contains majority of the drug targets that have inhibitors on the market.

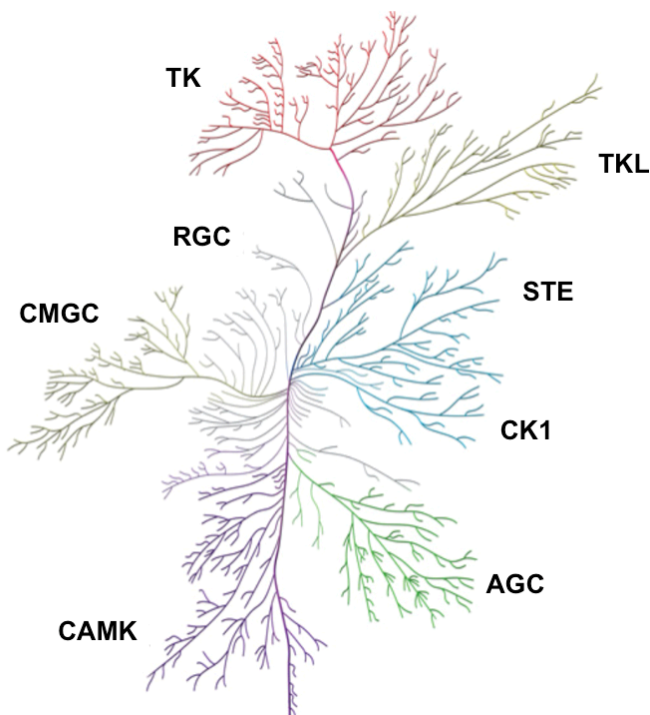


Figure 1.5: The eukaryotic protein kinase (ePK) protein kinome tree. Human ePK's are classified into 8 groups based on sequence similarity in the kinase domain, which are colored by group. Adapted from Duong-Ly and Peterson.⁵⁵

The AGC group is named after the enzyme families, protein kinase A (PKA), protein kinase G (PKG), and protein kinase C (PKC).⁷⁰ These enzymes are serine/threonine kinases regulated by cyclic adenosine monophosphate (cAMP) or lipids. A member within this group include AKT (PKB) kinases, which are particularly important for regulating cell growth, proliferation, protein synthesis, glucose metabolism, and survival.

The CMGC group is a diverse group of kinases. This includes the cyclin-dependent kinases (CDKs) and CDK-like kinases, which are central regulators of cell cycle progression. Since the second hallmark of cancer involves evasion of anti-proliferative signals (*discussed in Introduction to Cancer Section ii*), genetic or structural alteration of CDKs can promote tumorigenesis.¹ Also included in this group are the mitogen-activated protein kinases (MAPKs), which are involved in cell proliferation, differentiation, and apoptosis. Therefore, mutation of these enzymes will allow cancer cells to develop resistance towards apoptosis (*refer to third hallmark of cancer in Introduction to Cancer Section ii*). Lastly, the CMGC also consists of glycogen synthase kinases (GSKs), which are involved in inflammation and glycogen metabolism.

The TKL group consists mostly of serine/threonine kinases. They are named tyrosine kinase-like because they have sequences resembling those of the TK group. This group also contains receptor and non-receptor (cytosolic) protein kinases. Examples of protein kinases within this group include the interleukin1 (IL1) receptor-associated kinase and the transforming growth factor beta (TGF β) receptors.

The STE group comprises kinases that are homologous to the yeast proteins STE20, STE11, and STE7. Members within this group include the p21-activated kinases (Paks), which are critical regulators of diverse signaling pathways.

The CK1 group includes casein kinase 1 and homologous kinases. These protein kinases are also serine/threonine kinases that are constitutively expressed. They phosphorylate a diverse array of substrate molecules involved in cytoskeletal function and transcriptional regulation.

Kinases in the CAMK group are involved in calcium signaling and are basally auto-inhibited. Examples of kinases in this group includes the cell cycle checkpoint kinases, CHK1 and CHK2, which initiate a phosphorylation cascade leading to cell cycle arrest and repair of damaged DNA.

The smallest kinome group, RGC, is the receptor guanylyl cyclases. These kinases convert guanosine triphosphate (GTP) to cyclic guanine monophosphate (GMP). These kinases are termed pseudokinases because they lack certain residues that are critical for phosphate transfer.⁶⁷

Classification of kinases into the human kinome reveals how these enzymes are involved in a diverse array of cellular processes, most of which prevents transformation of normal cells to cancer cells. Therefore, protein kinases are commonly altered in some manner in various types of cancers, making protein kinases an important drug target.

(iv) Implication in Cancer

Due to their role in so many cellular processes, it is not surprising that abnormal phosphorylation can lead to the hallmarks of cancers,⁷¹⁻⁷³ cardiovascular diseases,⁷⁴ neurodegenerative diseases,⁷⁵ inflammatory diseases,^{76, 77} and diabetes.⁷⁸ The Cancer Gene Consensus (a literature-based consensus of genes that are mutated and causally implicated in cancer development) revealed that protein kinases were the most common protein domain encoded by cancer genes, with 27 of the 291 cancer genes encoding protein kinases.⁷⁹ Consistently, an analysis of the US Food and Drug Administration (FDA)-approved drugs since the 1980s indicated that kinases have surpassed GPCRs, a common important cancer drug target, as the most sought-after targets for cancer

treatment.⁵² To date, the U.S. Food and Drug Administration has approved 27 small molecule protein kinase inhibitors and 1 lipid kinase inhibitor.⁸⁰ These small molecule protein kinase inhibitors can be classified into two broad categories for inhibition type: ATP-competitive inhibitors and non-ATP-competitive inhibitors.

ATP-competitive inhibitors are kinase inhibitors whose potency depends on ATP concentrations, and are further classified as type I or type II inhibitors.⁸¹ Type I inhibitors bind to the protein kinase active DFG-in conformation whereas type II inhibitors bind the inactive DFG-out conformation.⁸² These inhibitors may contain a group that mimics the adenosine base of ATP, as seen in the type I inhibitors, erlotinib and gefitinib, which target EGFR. An example of a type II inhibitor is imatinib (*previously discussed in Introduction to Cancer Section iii*), which occupies the adenine pocket of the ATP binding site and a back hydrophobic pocket.

Non-ATP-competitive inhibitors potency does not vary with ATP concentration because they bind an allosteric site. Many of these allosteric inhibitors bind to regions outside of the kinase domain and regions unique to a particular kinase. Therefore, non-ATP-competitive inhibitors are often more selective than ATP-competitive inhibitors.

Although drug discovery for protein kinases has achieved a great deal of success, several significant challenges remain in the development of future drugs. First, evolutionary pressure results in the accumulation of point mutations in the kinase domain, which compromises inhibitor potency and leads to long-term drug resistance.⁸³ Second, the conserved architecture of the kinase domain within a class of protein kinases makes obtaining selectivity a challenge.^{76, 84} Third, the current kinase inhibitors on the market only covers a small subset of the human kinome, with 18 of the 27 approved

drugs covering only three out of more than 90 groups of tyrosine kinases, BCR-Abl, ErbBs, and VEGFRs. Given these shortcomings, and the importance of the target, there is a need to improve kinase drug discovery, where optimizing the enrichment of actives in virtual screening (VS) methods by using ensemble docking is one important avenue and one focus of this dissertation (*virtual screening will be discussed in more detail in Section iii of “MD Applied to Biological Macromolecules”*). In this work, six different protein kinases are used as a case study to determine the impact of using ensemble docking in enhancing VS performance against protein kinases.

(v) Protein Kinases in Study

Herein this dissertation, a case study of six protein kinases that span three kinome classes, CMGC, AGC, and TK, are utilized in benchmarking VS performance. We generate ensembles from molecular dynamics simulations in order to incorporate protein kinase dynamics into virtual screens. The utility of MD simulations in integrating protein dynamics into the drug discovery pipeline will be further discussed in *Section iii of Molecular Dynamics Applied to Biological Macromolecules*.

The mitogen-activated protein (MAP) kinase-activated protein kinase 2 (MK2; MAPKAP2) has emerged as a desirable target for safe anti-inflammatory drugs. MK2 belongs to the CMGC kinome family.⁶⁷ MK2 is one of several kinases directly phosphorylated and activated by the p38 MAP kinase. The activated MK2 activates substrates in both the nucleus,⁸⁵⁻⁸⁷ and cytoplasm.⁸⁸⁻⁹¹ Through these phosphorylation events in the nucleus and cytoplasm, MK2 is involved in several cellular processes

including stress and inflammatory responses, nuclear export, gene expression regulation, and cell proliferation.

Cyclin-dependent kinase 2 (CDK2) is also found within the CMGC kinome class.⁶⁷ CDKs are involved in cell cycle progression and transcription. CDK2 is a catalytic subunit in the CDK complex, whose formation is required to allow cells to progress from the G1 to S phase. Cyclin E binds CDK2, which allows cells to transition from G1 to S phase, and binding of cyclin A allows cells to progress through the S phase.^{4, 92} Since alterations in cell checkpoint regulation can lead to aberrant cell division, CDK2 is an attractive target for therapeutics designed to arrest or recover control of the cell cycle, such as cancer and Alzheimers disease.^{93, 94}

Rho-associated protein kinase 1 (ROCK1) is a member of the AGC kinome family.⁶⁷ ROCK1 is activated when GTP-bound RhoA⁹⁵ binds it, and is involved in cytoskeleton assembly and cell motility and contraction. Activated ROCK1 regulates the activity of muscle myosin regulatory light chain (MLC) protein via direct phosphorylation^{96, 97} and by phosphorylation and inhibition of the myosin binding subunit of myosin phosphatase. This in turn leads to increased levels of phosphorylated MLC and subsequent muscle contraction.⁹⁸ ROCK1 is also involved in nonmuscle myosin regulation and has been implicated in stress fiber and focal adhesion formation,⁹⁹ neurite retraction,¹⁰⁰ and tumor cell invasion.¹⁰¹ ROCK1 has several therapeutic indications including, cancer,^{102, 103} hypertension,¹⁰⁴ atherosclerosis,¹⁰⁵ and immunosuppression.¹⁰⁶

Protein kinase B (PKB or AKT) is also a member of the AGC kinome class.⁶⁷ AKT is a key player in the phosphoinositide3-kinase (PI3K) –AKT signaling pathway. AKT1 is activated via 3-phosphoinositide-dependent protein kinase 1 (PDK1)

phosphorylation at the plasma membrane. Activated AKT1 leads to activating a large number of substrates¹⁰⁷ involved in cell growth, proliferation, motility, and survival¹⁰⁸. The PI3K-AKT signaling pathway is one of the most frequently deregulated signaling pathways in human cancers and has been shown to mediate resistance of therapeutics.¹⁰⁹

Insulin-like growth factor-1 receptor (IGF-1R) is a member of the TK kinome class.⁶⁷ The IGF1R is a transmembrane receptor that is activated by the insulin-like growth factor 1 (IGF-1) hormone with high affinity, and a related hormone called IGF-2 and insulin with lower affinity. Activated IGF-1R activates several downstream cell-signaling cascades in the Ras/Raf/MAPK, PI3K/AKT pathways,¹¹⁰⁻¹¹⁶ and JAK/STAT pathway¹¹⁷. Activation of the MAPK pathway induces cellular proliferation, and PI3K/AKT pathway inhibits apoptosis and stimulates protein synthesis. Activation of the JAK/STAT pathway via phosphorylation of Janus kinases phosphorylates and activates signal transducers and activators of transcription (STAT) proteins.

The Abelson murine leukemia viral oncogene homolog 1 (ABL) non-receptor tyrosine kinase is found within the TK kinome class.⁶⁷ Phosphorylation of ABL1 via cell division cycle protein 2 (CDC2) allows ABL1 to bind DNA, suggesting a role in the cell cycle.¹¹⁸ Phosphorylation of nuclear and cytoplasmic substrates implicates ABL1 in cell differentiation, cell division, cell adhesion, and stress response.¹¹⁹ ABL1 contains an SH3 domain that negatively regulates its activity. Deletion of the SH3 domain turns ABL1 into an oncogene, a gene that has the potential to cause cancer.¹²⁰ For example, the ABL1 (deleted SH3) may fuse with the breakpoint cluster region (BCR), leading to a fusion gene, BCR-ABL1, which is present in many cases of chronic myelogenous leukemia.¹²¹

Tumor Suppressor, p53

(i) Biology of p53

The tumor suppressor p53 responds to several environmental stressors and induce either the expression or activation of proteins involved in stress response pathways (Figure 1.6). Under normal unstressed conditions, p53 exists in low concentrations through rapid ubiquitination and degradation via the E3 ubiquitin ligase, mouse double minute protein 2 (MDM2).¹²²⁻¹²⁶ A homolog of MDM2, MDM4, also serves as a negative regulator of p53.¹²⁷ Also, p53 mainly exists as a monomer (~30%) or dimer (~60%) under normal cellular conditions.¹²⁸

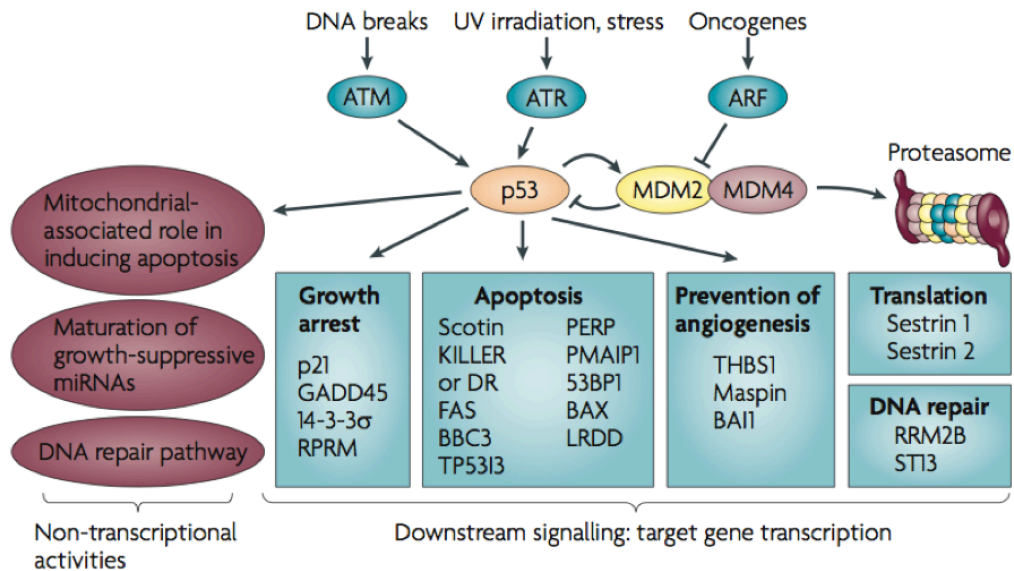


Figure 1.6: Summary of the p53 pathway.¹²⁶ p53 is at the center of the stress response pathway. Under normal unstressed cellular conditions, p53 is targeted for ubiquitination by MDM2 and MDM4. When the cell becomes stressed via DNA damage, UV radiation, or oncogene upregulation, the p53 complex with MDM2 and MDM4 is blocked, and p53 is upregulated. Upregulation of p53 leads to activation or expression of signaling molecules involved in stress in the stress response pathway, such as growth arrest, apoptosis, inhibition of angiogenesis, translation, and DNA repair.

When the cell experiences DNA damage, stress, or expression of oncogenes, the MDM2 interaction with p53 is disrupted (Figure 1.6). Several protein kinases are up

regulated and extensively phosphorylate the N-terminus of p53, which destabilizes p53 interaction with MDM2.¹²⁹⁻¹³¹ This blocks proteosomal degradation of p53, leading to up regulation of p53. This increase in concentration shifts p53 monomers and dimers into a tetrameric state (>90%).¹²⁸ p53 can then activate stress response signals through either non-transcriptional or transcriptional pathways (Figure 1.6).¹³²

While p53 can engage in direct protein-protein interactions in inducing stress response pathways, for example interactions between p53 and the apoptotic effector protein BAX;^{48, 133} majority of the p53-regulated stress responses occur through p53-directed activation of transcription. The genes activated by p53 range from those that activate apoptosis via p53 interactions with Bax and PUMA DNA response elements, to genes that induce senescence, cell cycle arrest, and DNA repair via interactions with p21, GADD45, PML, and PCNA response elements.^{48, 134} The negative regulator, MDM2, is also included in the DNA response genes activated via p53, whose function is to down regulate p53. The mechanism by which p53 searches and recognizes its response elements remains a topic of debate.^{135, 136} Studies suggest that in addition to the DNA binding domain, the C-terminal domain also plays a role in DNA search and recognition, which will be discussed in more detail next in *Section ii*.

(ii) Structure of p53

The full-length p53 protein (fl-p53) is a multi-domain, partially intrinsically disordered protein that binds DNA as a tetramer.¹³⁷ Fl-p53 consists of 393 amino acid residues that form a flexible N-terminal domain (NTD), a core DNA binding domain (DBD), a flexible linker region, a tetramerization domain (TET), and a flexible C-terminal domain (CTD) (Figure 1.7).

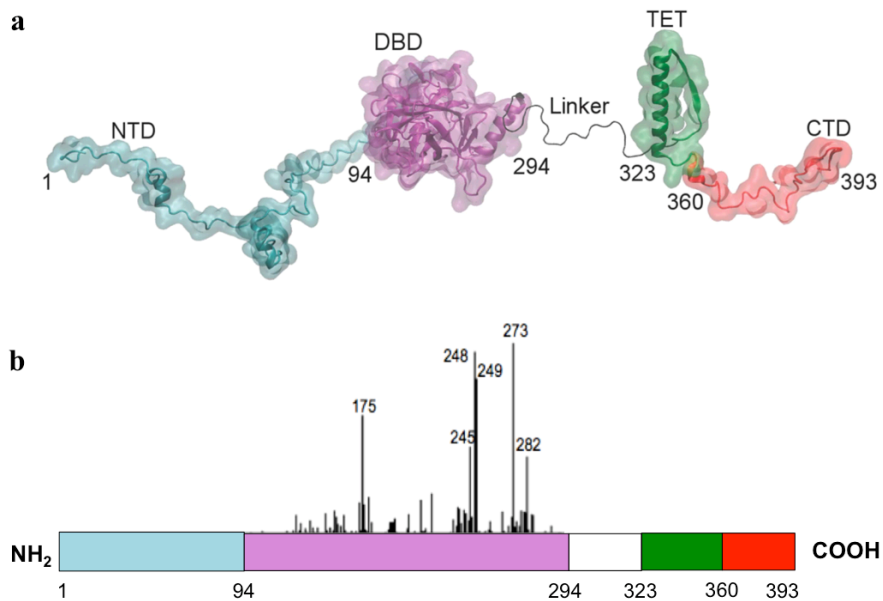


Figure 1.7: Structure of full-length p53. Each fl-p53 domain is shown in ribbon representation and colored cyan, purple, black, green, and red for the NTD, DBD, linker, TET, and CTD respectively (a). A simplified schematic representation of each domain is shown, with the mutation frequency within the DBD shown; the six hotspot mutations are labeled (b).

The first 94 residues comprise the highly dynamic NTD. Within the NTD, the transactivation domain (residues 1 to 61) is responsible for activating transcription factors and interacts with a wide variety of p53 targets.¹³⁸⁻¹⁴⁵ Specifically, the NTD contains two amphipathic subdomains, termed AD1 and AD2, where p53 target proteins bind in a ‘fly-casting’ mechanism¹⁴⁶ (describes how a disordered protein binds weakly and non-specifically to its target and folds as it approaches the cognate binding site).^{147, 148} Following the transactivation domain, the proline rich domain (PRD) comprises of residues 62 to 94. The PRD contains a series of PAAP repeats, where the amino acid composition is poorly conserved between species, but the overall length is conserved. Mouse model studies of p53 have suggested that the PRD plays a role in apoptosis.¹⁴⁹ The C-terminal region of the PRD (residues 90-94) forms stable intramolecular interactions with the DBD, which helps stabilize the p53 DBD and decrease aggregation propensity.¹⁵⁰

Following the NTD is the DBD (residues 95 to 294), which is the largest p53 domain, and has a defined secondary and tertiary structure. Due to this discrete fold within the DBD, there are many available crystal and NMR structures that reveal a well-defined β -sandwich fold that contains two large loops and a loop-sheet-helix motif.¹⁵⁰⁻¹⁵³ The DBD is the primary DNA interaction site of p53, binding DNA cooperatively to DNA response elements with a 4:1 p53:DNA stoichiometry.^{154, 155} Response elements are comprised of two 10-bp half-sites with the sequence, RRRCWWGYYY, where R is adenine or guanine, W is adenine or thymine, and Y is cytosine or thymine.¹⁵⁶ The fl-p53 tetramer is a dimer of dimers and each monomer binds to a pentamer repeat.¹⁵⁷

Within the DBD, sequence-specific DNA contacts are made through hydrophobic and electrostatic interactions. Residues in the L1 loop (Lys 120), S2, S2', and S10 β -strands, and the H2 helix contacts the major groove of DNA.^{147, 151, 152} Residues in the L3 loop (Ser 241 and R248) interact with the minor groove of the DNA, which is stabilized by a zinc ion that is situated between loops L2 and L3. While the L2 loop does not engage in direct DNA contacts, it stabilizes the L3 loop-DNA contacts via salt-bridges and the zinc ion, which is coordinated by residues from the L2 and L3 loops. Several X-ray crystal and electron microscopy (EM) structures have shown that the DBD tetramer binds on the same side of the DNA.

p53 binds specific response elements with low nanomolar affinity;^{152, 155, 158-160} however, binding affinities to non-specific DNA are an order of magnitude higher.^{152, 155} This suggests that the interactions between the DNA response elements and p53 DBD are crucial. Interestingly, the DBD contains majority of the p53 mutations found in human tumors (Figure 1.7).¹⁶¹ Therefore, the DBD holds the key to understanding how p53 binds DNA, and how tumorigenic mutations disrupt this DNA binding.

Located between the DBD and TET domains in p53 is a disordered linker region that comprises residues 295 to 322, and contains the dominant p53 nuclear localization signal (NLS).^{138, 162, 163} The NLS consists of residues 316-322 and 304-305.¹⁶³ Three Lys residues within the NLS (319-321) are ubiquitinated under normal unstressed cellular conditions, which inhibits p53 nuclear import.¹⁶⁴ When the cell becomes stressed, ubiquitination is blocked as a result of the disrupted p53:MDM2 interaction, which then allows p53 nuclear import.^{164, 165}

Following the linker region is the TET domain, which consists of residues 323-360, and contains a short N-terminal β -strand followed by a α -helix. The TET subunits associate in forming the tetrameric p53 state. Due to its stable fold, the TET domain has been well characterized by NMR and X-ray studies.¹⁶⁶⁻¹⁷¹ Each TET subunit assembles as a dimer of dimers through strong anti-parallel association of the β -strand and the α -helix, with a K_D of 500 pM.¹⁷² Both dimers further interact to form a tetramer of D2 symmetry through hydrophobic contacts, with a K_D of \sim 50nM.¹⁷²

The last 32 residues (361-393) of fl-p53 make up the CTD, a highly dynamic domain that engages in p53 target protein and DNA interactions.^{135, 147} Under stable cellular conditions, the CTD is ubiquitinated via p53 NTD:MDM2 interactions, inducing p53 down regulation and degradation.^{131, 144, 173} In response to the cellular stress response, disruption of the p53 NTD:MDM2 interactions leads to decreased CTD ubiquitination, leading to up regulation of p53 and activation of proteins involved in stress pathways. During p53-mediated transcriptional activation, the CTD is involved in extensive DNA interactions.^{147, 174-176}

While experimental studies have shown that the CTD plays a role in p53-DNA binding, the exact role of the CTD has remained controversial.¹⁷⁶ Initial studies suggested that the CTD acts as a negative regulator by blocking DBD tetramers from binding DNA.¹⁷⁷⁻¹⁸⁰ Contrarily, further research suggested that the CTD promotes binding to both linear and non-linear DNA, suggesting a positive regulator role for the CTD.^{135, 159, 181-186} One study proposed a search and recognition binding mode for DNA, in which the CTDs facilitate target search by sliding along the DNA making non-specific interactions, while the DBDs engage in frequent association and dissociation.^{181, 187, 188} In

addition, computational studies of the fl-p53 bound to three different DNA response elements revealed that positively-charged residues (Lys and Arg) within the CTD approached and directly contacted the DNA independent of the response element.¹⁷⁴ Interestingly, experimental studies where Lys residues within the CTD (372, 373, 381, 382) were acetylated showed decreased DNA binding by the isolated CTD in vitro,¹⁷⁵ providing further evidence for CTD's positive regulation in p53 DNA binding.

(iii) Implication in Cancer

With more than a thousand p53 mutations in human tumors, it is now widely accepted that p53 mutations are the most common genetic event in human cancer.⁴⁸⁻⁵¹ The majority of these p53 mutations (>90%) occur in the DBD in which there exist six hotspot mutations (R175, G245, R248, R249, R273, and R282) that occur at an unusual high frequency (Figure 1.7).^{151, 189, 190} While wildtype p53 has a short-lived half-life, mutant p53 has a pro-longed half-life.¹⁹¹ The effect of each mutant is different, and why tumors select for one mutation over another is still unclear. However, p53 mutants are broadly categorized into two classes.¹⁹² One class of mutations is contact mutants, which involves residues that directly contact DNA, and loss of the contact leads to disrupted DNA binding. The other class of mutations is structural mutants, which includes residues that are important for the stable folding of the DBD, and loss of DNA binding is due to structural defects.

The best example of a structural mutant and the focus of this dissertation is the R175 mutation because R175 plays a critical role in stabilizing loops L2 and L3, which contain residues that make crucial DNA contacts. Several experimental studies suggest

that the R175 mutants are unfolded. One study found that R175 mutants associate with heat shock protein, hsp70, suggesting their partial denaturation.¹⁹³ Also, R175 mutants bind an antibody, Pab240, which recognizes mutant p53 or denatured p53 and not wildtype p53.¹⁹⁴ The mutant epitopes within the DBD are buried in wildtype p53; therefore, denatured p53 recognizes the Pab240 antibody. Third, R175 mutants are very sensitive to proteolytic enzymes, unlike wildtype p53.^{195, 196} Lastly, fusion proteins containing the full-length R175 mutant and DBD of GAL4 does not activate transcription, suggesting long-range denaturation effects in the NTD transactivation domain.¹⁹⁷

The actual mechanism of the pro-oncogenic effects of p53 mutants may vary, in which three are proposed.¹⁹⁸ First, tumors may select for p53 mutations that solely result in the loss of p53 tumor-suppressive functions. Second, p53 mutations may result in the loss of certain p53 tumor suppressive functions, while retaining or exaggerating other aspects of wildtype p53 function. Lastly, p53 mutants may acquire novel p53 functions that specifically promote tumorigenesis; this neomorphic activity describes p53 mutations gain-of-function abilities. Mutant p53 can acquire additional functions to promote cancer progression through both non-transcriptional and transcriptional interactions.

Mutant p53 can form aberrant protein complexes with several proteins. The most widely studied mutant p53-interacting partners include the p53 family members, p63 and p73. Several studies showed that mutant p53 forms heterotetramers with p63 and p73,¹⁹⁹⁻²⁰³ and this heterotetramer formation has been linked to promotion of chemoresistance, migration, invasion, and metastasis.²⁰⁴⁻²⁰⁶ This was quite surprising as wildtype p53 is

unable to form heterotetramers with neither p63 nor p73,²⁰⁷ suggesting that the gain-of-function of mutant p53 leads to inhibition of p63 and p73.

In addition to non-transcriptional mediated gain-of-function activities, mutant p53 has also been shown to transactivate genes involved in many different aspects of tumorigenesis.^{191, 198, 208} For example, mutant p53 can transactivate genes that promote proliferation of cancer cells, IGF-1R for example (From Freed-Pastor paper: Werner et al. 1996). Mutant p53 can also upregulate genes that inhibit apoptosis or promote chemoresistance,^{204, 209-218} all of which can inhibit cell death of cancer cells.

(iv) p53 Drug discovery

In vivo studies have shown that reactivation of wildtype p53 in p53-null or p53 mutant tumors are sufficient to regress tumor progression.²¹⁹⁻²²³ Therefore, multiple approaches to restore wildtype p53 in tumor cells are being utilized in developing therapies. One strategy involves developing small molecules that block p53 interaction with its negative regulators (MDM2 for example) and block the activity of cellular factors that inhibit wildtype functionality, leading to upregulation of wildtype p53 in tumors.²²⁴⁻²²⁷ Another approach includes gene therapy, where wildtype p53 is delivered to tumors.²²⁴⁻²²⁷ An alternative therapeutic approach includes small molecules that specifically reactivate mutant p53 to wildtype conformation or destabilize mutant p53.

The first example of a small molecule that specifically targeted mutant p53 was CP31398, which induces expression of canonical p53 target genes or drive expression from a p53 reporter construct in cells expressing mutant p53, impairing tumor growth.²²⁸ However, this molecule was later found to intercalate DNA instead of binding mutant

p53.²²⁹ The most well advanced example of a small molecule that specifically targeted mutant p53 is PRIMA-1, which reactivates missense mutations of p53 which regains some wildtype functions of p53 and halt tumor growth.^{230, 231} PRIMA-1 is currently in clinical trials as a treatment for ovarian and myeloma cancers.²³²

In addition to binding mutant p53 in order to reactivate wildtype p53 function, another approach to targeting p53 in human cancers involves designing small molecules to destabilize mutant p53 gain-of-function conformations. One example of destabilizing mutant p53 is to disrupt the p53:p63/p73 interaction, which has been implicated in many pro-oncogenic effects of mutant p53. An example of such a small molecule is RETRA (reactivation of transcriptional reporter activity), where studies have shown that it blocks mutant p53 interaction with p73, and prevents xenografted tumor cell growth.²³³ Another approach in destabilizing p53 mutants involves inhibiting factors that function to stabilize mutant p53 in tumors, such as HDAC6 or Hsp60.^{234, 235}

The last therapeutic approach that will be discussed and applies to this dissertation work involves developing a structure-based ‘mutant-specific’ drug. For example, PhiKan083 is small molecule that binds the C-terminus of the Y220C p53 mutant, which stabilizes the DBD and restores transactivation of p53 target genes.²³⁶ Another example involves Stictic acid, which has been shown to reactivate the R175H p53 mutant in both *in vitro* and *in vivo* studies.²³⁷ The utility of these small molecules as actual drugs are still being explored. However, both provide excellent proof of principles for rational drug discovery targeting specific p53 mutants.

While there are multiple avenues being explored in developing therapies for p53, there still exist major challenges. For starters, there are not experimental structures for

most p53 mutants, especially structural mutants. In the case of the Y220C mutant, there are multiple crystal structures of the Y220C p53 mutant, which was used in the design of PhiKan083 as discussed previously. However, for other mutants such as R175H, no experimental structures are available due to the denaturation effects of this mutant. Although experiments do in fact show reactivation of the R175H mutant to wildtype after Stictic acid binding, we are not entirely clear if and how Stictic acid binds the R175H p53 mutant. Second, we don't fully understand the dynamic behavior of p53 mutations and how their dynamic behaviors differ from wildtype p53. Delineating the dynamic characteristics of both wildtype and mutant p53 will allow us to target specific conformational states explored by the mutant and not wildtype with drug molecules. Third, we still do not understand the p53-DNA binding mechanism under normal wildtype p53 conditions at a molecular level. Therefore, we do not understand how this DNA binding mechanism is disrupted as a result of p53 mutations. Computational studies have revealed a clamping and symmetric quaternary binding mode of the DBD's when binding DNA response elements.¹⁷⁴ In this dissertation, I aim to understand how p53 cancer mutations, R175H specifically, alters this DNA binding mode in aiding in the development of future p53 reactivation molecules. Molecular dynamics is used to explore and compare wildtype and R175H mutant fl-p53 dynamics. The theory and application of molecular dynamics is discussed in more detail next.

Molecular Dynamics Applied to Biological Macromolecules

(i) Protein Dynamics

Proteins do not function as static systems, but are dynamic.²³⁸⁻²⁴¹ In fact, Weber characterized proteins as “screaming and kicking”.²⁴² Protein dynamics are represented as an ensemble in an energy landscape, which describes the potential energy of a protein as a function of the conformational coordinates of a protein (Figure 1.8).²⁴³ Each basin represents a ‘conformational substate’ or ‘microstate’ that the protein hops between at any given moment. Protein motions can be defined as transitions between these microstates, and the energy barrier between different microstates determines the transition rate between microstates.

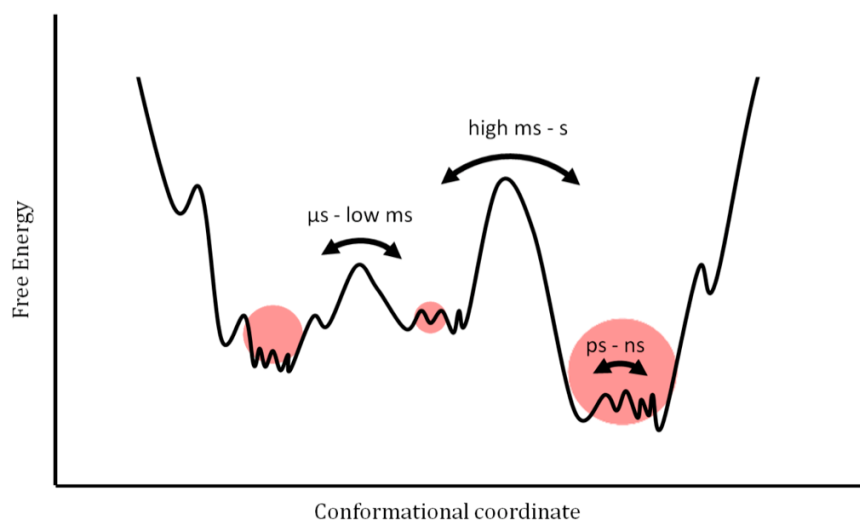


Figure 1.8: Example free-energy landscape of proteins. Each basin represents different microstates, where minima correspond to well-defined, stable states, and maxima reflect short-lived transition states. The height of the activation energy barriers is proportional to the timescale that is necessary to transition between microstates. The volume of the red spheres depicts the hypothetical population of conformers. Extracted from Göbl and Tjandra.²⁴¹

Protein motions are directly related to the function of proteins, permitting ligands to bind a protein, substrates to reach the active center of enzymes, and catalytic groups to come together.²⁴⁴⁻²⁴⁷ Chemical modifications and mutations of proteins can affect the populations of their ensemble.²⁴⁸⁻²⁵¹ Many experimental techniques allow us to study protein motions, where X-ray crystallography and nuclear magnetic resonance (NMR) provide information about protein fluctuations with atomistic resolution.

X-ray crystallography is commonly used for atomistic structure determination of well-folded biomolecules. Dynamic information about a protein can be provided via protein crystallization of two endpoints of a reaction, such as the unbound and ligand-bound state of a protein. The crystals of both protein states would allow a linear interpretation of rearrangements that take place during ligand binding. While X-ray crystallography provides atomic level resolution of protein structures, where dynamic information can be inferred, nuclear magnetic resonance (NMR) spectroscopy reveals both the structure and dynamics of proteins. NMR measures both internal distances between atoms and atomic motions in a protein using the nuclear spin magnetic moments of ^1H , ^{13}C , and ^{15}N , atoms that make up the backbone of proteins. Unlike X-ray crystallography, an NMR structure of a protein contains multiple frames or snapshots of a protein that provides information about the dynamic characteristics of the protein.

X-ray crystallography and NMR may reveal insight into how structural alterations of a protein, such as a cancer mutation, impact the structure and dynamics of a protein. For example, protein crystallization and NMR can be used to resolve the normal wildtype and mutant state of a protein. Comparison of these structures may reveal how the mutation alters the structure and impede the function of the protein. However, the

changes may be too subtle or no structural change may occur at all. In addition, X-ray crystallography and NMR studies may not be sufficient in resolving the structures if the protein is intrinsically disordered or if the mutation denatures the protein (*recall R175H mutation as discussed previously*). Also, the process of generating these structures is expensive and labor intensive. Therefore, the use of molecular dynamics is an attractive alternative as it is the only method where the structure and dynamics can be studied simultaneously at atomistic resolution.

(ii) Theory of Molecular Dynamics

Molecular dynamics (MD) simulations provide a dynamic evolution of atoms within a protein through the use of molecular mechanics force fields. In molecular mechanics, the atoms are represented as nodes in space based on their Cartesian coordinates in the x-, y-, and z- direction, and bonds are represented as edges connecting the nodes.²⁵² The nodes are described based on their atom type and hybridization, and the length of the edges is based on the bond lengths. The reason we are able to apply molecular mechanics or empirical force fields to large systems, such as proteins, is due to the validity of several assumptions. The first assumption is the Born-Oppenheimer approximation, which allows us to separate the electronic and nuclear motions. Therefore, we are able to simplify molecular motions calculations in MD simulations by ignoring the electronic motions, the focus of quantum mechanical models, and calculate the energy of the system as a function of the nuclear positions only. The result is a simplified model of the interactions within the system with contributions from bond

stretching, opening and closing of angles, rotations around single bonds, and long-range atomic interactions.²⁵³

The force fields used in MD simulations can be described as a relatively simple four-component picture of intra- and inter-molecular forces within the system (Figure 1.9).^{254, 255}

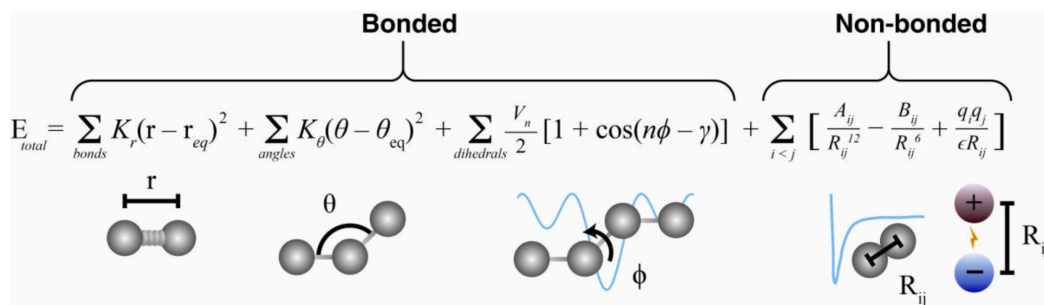


Figure 1.9: Example force field equation used in molecular dynamics simulations. The total potential energy can be divided into both the bonded and non-bonded interactions. The chemical bonds and atomic angles are modeled using simple springs (first and second terms), and dihedral angles are modeled using a sinusoidal function that approximates energy differences between eclipsed and staggered conformations (third term). The van der Waals interactions are modeled using the Lennard-Jones potential, and electrostatic interactions are modeled using Coulomb's law. Extracted from Durrant and McCammon.²⁵¹

E_{total} is the total potential energy, which is a function of the atomic positions (r) of N atoms. The first term in the equation shown in Figure 1.9 sums over all the interactions between bonded atoms, which is modeled by a harmonic potential that measures the increase in energy as the bond length, r , deviates from the reference bond length, r_{eq} . The second term is a summation over all angles also modeled using a harmonic potential, where θ is the deviation from the reference angle, θ_{eq} . In the bonds and angles summations, K_r and K_θ are the spring constants. The third term is a summation over all the dihedral angles, where V_n is a weighting factor used in measuring the deviation, ϕ ,

from the reference dihedral angle, γ . The fourth contribution to the potential energy function includes the non-bonded atomic interactions. The Lennard-Jones potential²⁵⁶ is used to model the van der Waals interaction, where A and B are experimentally determined values, and R_{ij} is the distance between atoms. The Coulomb potential term measures the electrostatic potential, which is proportional to the charges of the two atoms, $q_i q_j$, and inversely proportional to the distance between the atoms R_{ij} . Several force fields are used in MD simulations such as, AMBER,^{254, 257} CHARMM,²⁵⁸ and GROMOS²⁵⁹. While different force fields used in MD simulations all use the general form described in equation 1, there are different parameters and weighting factors used, and generally yield similar results.²⁶⁰

There remain two main limitations associated with MD simulations: (i) the amount of simulation time accessible with available computer resources and sampling algorithms (sampling problem), and (ii) the accuracy of the potential energy function (force field problem). The development of enhanced MD sampling methods such as accelerated MD, metadynamics, temperature enhanced MD, replica exchange MD, and more²⁶¹ allows us to simulate more phase space than conventional MD methods. Also, the design of computational hardware for speeding up simulation calculations is the most effective means in increasing the sampling limit of MD simulations. An example of this includes the use of Graphics Processing Units (GPUs), which were used to perform MD simulations in this dissertation. Interestingly, the advances in the sampling problem highlighted the inaccuracies in force fields, as many force field defects only became apparent with longer simulations.²⁶²⁻²⁷¹

These findings have led to a series of reparametrizations in all the major protein force fields. For example, the parameters for the ϕ - and ψ - torsion angles were altered in the OPLS-AA/L²⁷² and AMBER force field^{268, 273}, whereas in the CHARMM force field, a grid-based energy correction for the ϕ - ψ plane was introduced^{274, 275}. More recently, updated parameters for the χ_1 torsion angle were published for the AMBER and CHARMM force fields^{271, 276-278}, the two most commonly used force fields.^{273, 279} These reparametrized force fields show significant improved agreement with experimental data.^{277, 280, 281}

(iii) MD Simulations Utility in Drug Discovery

With constant improvements in the sampling problem and MD force fields, MD simulations are likely to play an increasingly important role in the drug discovery pipeline. For starters, MD simulations may identify cryptic or allosteric binding sites that experimental structures may not capture. For example, Schames *et al.* performed MD simulations on HIV integrase,²⁸² an enzyme produced by HIV that enables its genetic material to be integrated into the DNA of the infected cell. The simulations revealed a novel pocket that was not evident from available X-ray crystal structures. Later, X-ray crystal structures demonstrated that known HIV integrase inhibitors do in fact bind this cryptic site. These results led to experimental studies at Merck & Co²⁸³, where further development yielded production of the first US FDA-approved highly effective antiretroviral drug raltegravir.²⁸⁴

Another example of identification of a cryptic pocket from MD that served as the foundation for this dissertation involves the tumor suppressor p53. Wassman and

coworkers performed MD simulations of wildtype and various cancer mutants of p53.²³⁷ The simulations revealed a transiently open binding pocket in the DNA binding domain between loop L1 (a loop that makes crucial DNA contacts) and beta strand 3. Virtual screening against this novel pocket and *in vivo* experimental studies yielded Stictic acid as a potential p53 reactivation compound. These promising results contributed to the founding of the company, Actavalon.²⁸⁵

In addition to identification of cryptic binding pockets, MD can enhance *in silico* traditional VS methods. A docking program is used to predict the binding pose and binding affinity of small molecules within a selected receptor-binding pocket. Typically, ligand databases of compounds that are commercially available and synthetically accessible are docked into a single static receptor structure, as determined from NMR or X-ray crystallography. The best predicted ligands are selected for further experimental testing. Unfortunately, traditional docking neglects the dynamic characteristics of receptors. Some small molecule ligands may in fact bind the single receptor structure selected, but in reality receptors have many conformational states, where any of them may be druggable. Therefore, true ligands (that may be potential drug candidates) are often discarded because they bind a receptor conformation different from that of the single static structure chosen. In order to better accommodate receptor flexibility in virtual screens, a new VS protocol has been developed called the relaxed complex scheme (RSC).^{286, 287}

In the RSC protocol, each ligand is docked into multiple protein conformations typically extracted from MD simulations as opposed to docking into a single static structure. Thus, a range of docking scores is assigned to each ligand instead of one single

score. Ligands can then be ranked by several characteristics, such as the average score over all receptor conformations. Alternatively, the range of docking scores can be used to train receptor conformations as performed in this dissertation using the method, Ensemble Builder.²⁸⁸ RSC has been successful in identifying a number of protein inhibitors, including inhibitors of FK506 binding proteins,²⁸⁹ HIV integrase,²⁸² *Trypanosoma brucei* RNA editing ligase 1,^{290, 291} *T. brucei* GalE,²⁹² *T. brucei* FPPS,²⁹³ *Mycobacterium tuberculosis* dTDP-6-deoxy-L-lyxo-4-hexulose,²⁹⁴ and p53²³⁷. While these successes are promising, the RSC protocol relies on docking scoring functions that are optimized for speed at the expense of accuracy. These scoring functions do not accurately account for conformational entropy and solvation energy in binding energies,^{295, 296} thereby sacrificing accuracy in predicting binding affinities. One way to overcome the limitations of docking scoring functions is through advanced free energy calculations using MD.

Although they are computationally expensive, techniques for predicting binding affinities more accurately do exist. These techniques include thermodynamic integration,²⁶¹ single-step perturbation,²⁹⁷ and free energy perturbation²⁹⁸. Since free energy is a state function, the free energy depends only on the initial energy in solution and the final energy following the binding event. The path of ligand binding only influences receptor-ligand kinetics, but it has no bearing on the free energy. Therefore simulating an entire ligand-binding event is not necessary in obtaining the free energy. Instead, a drug's binding affinity is calculated using a technique called 'alchemical transformation'.²⁹⁹ During the MD simulation, the electrostatic and van der Waals produced by ligand atoms are turned down gradually, eventually annihilating the ligand

from the receptor-binding site. The successful application of alchemical techniques has made these techniques promising in accurately predicting binding affinities.³⁰⁰⁻³⁰⁶ However, it is important to highlight that alchemical techniques are uniquely sensitive to inadequate conformational sampling.³⁰⁷ If MD simulations fail to sample system conformations *in silico* that are in fact sampled under biological conditions, predicted binding affinities will be incorrect.

This is not the case when inadequate MD sampling is used to identify cryptic pockets, allosteric sites, or pharmacologically relevant binding pocket conformations for VS. In these cases, some suitable receptor conformations may be missed; however, the conformations that are identified are still useful. Therefore, the results of the simulations are therefore incomplete, but not necessarily wrong. Short timescale MD simulations in this dissertation are used in two applications: (i) to explore and compare the conformational dynamics of p53 under normal wildtype and mutant conditions, and (ii) to enhance VS performance against protein kinases.

References

1. Weinberg, D. H. a. R. A., The Hallmarks of Cancer. *Cell* **2000**, 100, 57-70.
2. National Cancer Institute, Cancer Statistics; <http://www.cancer.gov/about-cancer/understanding/statistics>.
3. Medema, R. H.; Bos, J. L., The role of p21ras in receptor tyrosine kinase signaling. *Crit Rev Oncog* **1993**, 4, 615-661.
4. Weinberg, R. A., The retinoblastoma protein and cell cycle control. *Cell* **1995**, 81, 323-330.
5. Datto, M. B.; Hu, P. P.; Kowalik, T. F.; Yingling, J.; Wang, X. F., The viral oncoprotein E1A blocks transforming growth factor beta-mediated induction of p21/WAF1/Cip1 and p15/INK4B. *Mol Cell Biol* **1997**, 17, 2030-2037.
6. Hannon, G. J. a. B., D., P15INK4B is a potential effector of TGF-beta-induced cell cycle arrest. *Nature* **1994**, 371, 257-261.
7. Fynan, T. M.; Reiss, M., Resistance to inhibition of cell growth by transforming growth factor-beta and its role in oncogenesis. *Crit Rev Oncog* **1993**, 4, 493-540.
8. Schutte, M.; Hruban, R. H.; Hedrick, L.; Cho, K. R.; Nadasdy, G. M.; Weinstein, C. L.; Bova, G. S.; Isaacs, W. B.; Cairns, P.; Nawroz, H.; Sidransky, D.; Casero, R. A., Jr.; Meltzer, P. S.; Hahn, S. A.; Kern, S. E., DPC4 gene in various tumor types. *Cancer Res* **1996**, 56, 2527-2530.
9. Chin, L., Pomerantz, J., and DePinho, R.A., The INK4/ARF tumor suppressor: one gene-two products-two pathways. *Trends Biochem Sci* **1998**, 23, 291-296.
10. Zuo, L., Weger, J., Yang, Q., Goldstein, A.M., Tucker, M.A., Walker, G.J., Hayward, N., and Dracopoli, N.C., Germline mutations in the p16INK4A binding domain of CDK4 in familial melanoma. *Nat Genet* **1996**, 12, 97-99.
11. Dyson, N., Howley, P.M., Munger, K., and Harlow, E., The human papillomavirus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science* **1989**, 243, 934-937.
12. Wyllie, A. H.; Kerr, J. F.; Currie, A. R., Cell death: the significance of apoptosis. *Int Rev Cytol* **1980**, 68, 251-306.
13. Butt, A. J.; Firth, S. M.; Baxter, R. C., The IGF axis and programmed cell death. *Immunol Cell Biol* **1999**, 77, 256-262.

14. Lotem, J.; Sachs, L., Control of apoptosis in hematopoiesis and leukemia by cytokines, tumor suppressor and oncogenes. *Leukemia* **1996**, 10, 925-931.
15. Green, D. R.; Reed, J. C., Mitochondria and apoptosis. *Science* **1998**, 281, 1309-1312.
16. Hayflick, L., Mortality and immortality at the cellular level. A review. *Biochemistry Mosc* **1997**, 62, 1180-1190.
17. Wright, W. E.; Pereira-Smith, O. M.; Shay, J. W., Reversible cellular senescence: implications for immortalization of normal human diploid fibroblasts. *Mol Cell Biol* **1989**, 9, 3088-3092.
18. Counter, C. M.; Ailion, A. A.; LeFeuvre, C. E.; Stewart, N. G.; Greider, C. W.; Harley, C. B.; Bacchetti, S., Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J* **1992**, 11, 1921-1929.
19. Shay, J. W.; Bacchetti, S., A survey of telomerase activity in human cancer. *Eur J Cancer* **1997**, 33, 787-791.
20. Bryan, T. M.; Cech, T. R., Telomerase and the maintenance of chromosome ends. *Curr Opin Cell Biol* **1999**, 11, 318-324.
21. Bryan, T. M.; Englezou, A.; Gupta, J.; Bacchetti, S.; Reddel, R. R., Telomere elongation in immortal human cells without detectable telomerase activity. *EMBO J* **1995**, 14, 4240-4248.
22. Veikkola, T.; Alitalo, K., VEGFs, receptors and angiogenesis. *Sem Cancer Biol* **1999**, 9, 211-220.
23. Fedi, P., Tronick, S.R., and Aaronson, S.A., *Growth Factors. In Cancer Medicine.* Williams and Wilkins: Baltimore, MD, 1997.
24. Varner, J. A.; Cheresch, D. A., Integrins and cancer. *Curr Opin Cell Biol* **1996**, 8, 724-730.
25. Giancotti, F. G.; Ruoslahti, E., Integrin signaling. *Science* **1999**, 285, 1028-1032.
26. Stetler-Stevenson, W. G., Matrix metalloproteinases in angiogenesis: a moving target for therapeutic intervention. *J Clin Invest* **1999**, 103, 1237-1241.

27. Dameron, K. M.; Volpert, O. V.; Tainsky, M. A.; Bouck, N., Control of angiogenesis in fibroblasts by p53 regulation of thrombospondin-1. *Science* **1994**, 265, 1582-1584.
28. Sporn, M. B., The war on cancer. *Lancet* **1996**, 347, 1377-1381.
29. Aplin, A. E.; Howe, A.; Alahari, S. K.; Juliano, R. L., Signal transduction and signal modulation by cell adhesion receptors: the role of integrins, cadherins, immunoglobulin-cell adhesion molecules, and selectins. *Pharmacol Rev* **1998**, 50, 197-263.
30. Christofori, G.; Semb, H., The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends Biochem Sci* **1999**, 24, 73-76.
31. Johnson, J. P., Cell adhesion molecules of the immunoglobulin supergene family and their role in malignant transformation and progression to metastatic disease. *Cancer Metastasis Rev* **1991**, 10, 11-22.
32. Kaiser, U.; Auerbach, B.; Oldenburg, M., The neural cell adhesion molecule NCAM in multiple myeloma. *Leuk Lymphoma* **1996**, 20, 389-395.
33. Fogar, P.; Basso, D.; Pasquali, C.; De Paoli, M.; Sperti, C.; Roveroni, G.; Pedrazzoli, S.; Plebani, M., Neural cell adhesion molecule (N-CAM) in gastrointestinal neoplasias. *Anticancer Res* **1997**, 17, 1227-1230.
34. Perl, A. K.; Dahl, U.; Wilgenbus, P.; Cremer, H.; Semb, H.; Christofori, G., Reduced expression of neural cell adhesion molecule induces metastatic dissemination of pancreatic beta tumor cells. *Nat Med* **1999**, 5, 286-291.
35. Lukashev, M. E.; Werb, Z., ECM signalling: orchestrating cell behaviour and misbehaviour. *Trends Cell Biol* **1998**, 8, 437-441.
36. Coussens, L. M.; Werb, Z., Matrix metalloproteinases and the development of cancer. *Chem Biol* **1996**, 3, 895-904.
37. Chambers, A. F.; Matrisian, L. M., Changing views of the role of matrix metalloproteinases in metastasis. *Journal Natl Cancer Inst* **1997**, 89, 1260-1270.
38. Werb, Z., ECM and cell surface proteolysis: regulating cellular ecology. *Cell* **1997**, 91, 439-442.
39. Bergers, G., and Coussens, L.M., Extrinsic regulators of epithelial tumor progression: metalloproteinases. *Curr Opin Genet Dev* **2000**, 10, 120-127.

40. National Cancer Institute, Types of Cancer Treatment, <http://www.cancer.gov/about-cancer/treatment/types>.
41. Urruticoechea, A.; Alemany, R.; Balart, J.; Villanueva, A.; Vinals, F.; Capella, G., Recent advances in cancer therapy: an overview. *Current Pharm Des* **2010**, 16, 3-10.
42. Lind, M. J., Principles of cytotoxic chemotherapy. *Medicine* **2011**, 39, 711-716.
43. Beatson, G., On the treatment of inoperable cases of carcinoma of the mamma: Suggestions for a new method of treatment, with illustrative cases. *Lancet* **1896**, 2, 104-107.
44. Smith, I. E.; Dowsett, M., Aromatase inhibitors in breast cancer. *New Engl J Med* **2003**, 348, 2431-2442.
45. Yarden, Y.; Sliwkowski, M. X., Untangling the ErbB signalling network. *Nat Rev Mol Cell Biol* **2001**, 2, 127-137.
46. Slamon, D. J.; Godolphin, W.; Jones, L. A.; Holt, J. A.; Wong, S. G.; Keith, D. E.; Levin, W. J.; Stuart, S. G.; Udove, J.; Ullrich, A.; et al., Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* **1989**, 244, 707-712.
47. Lydon, N. B.; Druker, B. J., Lessons learned from the development of imatinib. *Leuk Res* **2004**, 28 Suppl 1, S29-38.
48. Levine, A. J.; Oren, M., The first 30 years of p53: growing ever more complex. *Nat Rev Cancer* **2009**, 9, 749-758.
49. Hollstein, M.; Sidransky, D.; Vogelstein, B.; Harris, C. C., p53 mutations in human cancers. *Science* **1991**, 253, 49-53.
50. Nigro, J. M.; Baker, S. J.; Preisinger, A. C.; Jessup, J. M.; Hostetter, R.; Cleary, K.; Bigner, S. H.; Davidson, N.; Baylin, S.; Devilee, P.; et al., Mutations in the p53 gene occur in diverse human tumour types. *Nature* **1989**, 342, 705-708.
51. Takahashi, T.; Nau, M. M.; Chiba, I.; Birrer, M. J.; Rosenberg, R. K.; Vinocour, M.; Levitt, M.; Pass, H.; Gazdar, A. F.; Minna, J. D., p53: a frequent target for genetic abnormalities in lung cancer. *Science* **1989**, 246, 491-494.
52. Kinch, M. S., An analysis of FDA-approved drugs for oncology. *Drug Discov Today* **2014**, 19, 1831-1835.

53. Johnson, L. N.; Lewis, R. J., Structural basis for control by phosphorylation. *Chem Rev* **2001**, 101, 2209-2242.
54. Hanks, S. K.; Quinn, A. M.; Hunter, T., The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **1988**, 241, 42-52.
55. Adams, J. A., Kinetic and catalytic mechanisms of protein kinases. *Chem Rev* **2001**, 101, 2271-2290.
56. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, Y.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglu, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.;

- Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kasprzyk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowki, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J.; Szustakowki, J.; International Human Genome Sequencing, C., Initial sequencing and analysis of the human genome. *Nature* **2001**, 409, 860-921.
57. Rubin, G. M.; Yandell, M. D.; Wortman, J. R.; Gabor Miklos, G. L.; Nelson, C. R.; Hariharan, I. K.; Fortini, M. E.; Li, P. W.; Apweiler, R.; Fleischmann, W.; Cherry, J. M.; Henikoff, S.; Skupski, M. P.; Misra, S.; Ashburner, M.; Birney, E.; Boguski, M. S.; Brody, T.; Brokstein, P.; Celniker, S. E.; Chervitz, S. A.; Coates, D.; Cravchik, A.; Gabrielian, A.; Galle, R. F.; Gelbart, W. M.; George, R. A.; Goldstein, L. S.; Gong, F.; Guan, P.; Harris, N. L.; Hay, B. A.; Hoskins, R. A.; Li, J.; Li, Z.; Hynes, R. O.; Jones, S. J.; Kuehl, P. M.; Lemaître, B.; Littleton, J. T.; Morrison, D. K.; Mungall, C.; O'Farrell, P. H.; Pickeral, O. K.; Shue, C.; Vossell, L. B.; Zhang, J.; Zhao, Q.; Zheng, X. H.; Lewis, S., Comparative genomics of the eukaryotes. *Science* **2000**, 287, 2204-2215.
58. Manning, G.; Plowman, G. D.; Hunter, T.; Sudarsanam, S., Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* **2002**, 27, 514-520.
59. Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A. A.;

Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigo, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X., The sequence of the human genome. *Science* **2001**, 291, 1304-1351.

60. Hunter, T.; Plowman, G. D., The protein kinases of budding yeast: six score and more. *Trends Biochem Sci* **1997**, 22, 18-22.
61. Duong-Ly, K. C.; Peterson, J. R., The human kinome and kinase inhibition. *Curr Protoc Pharmacol* **2013**, Chapter 2, Unit2 9.
62. Ducommun, B.; Brambilla, P.; Felix, M. A.; Franza, B. R., Jr.; Karsenti, E.; Draetta, G., cdc2 phosphorylation is required for its interaction with cyclin. *EMBO J* **1991**, 10, 3311-3319.
63. Bossemeyer, D., The glycine-rich sequence of protein kinases: a multifunctional element. *Trends Biochem Sci* **1994**, 19, 201-205.
64. Bossemeyer, D., Loss of kinase activity. *Nature* **1993**, 363, 590.
65. Vetrie, D.; Vorechovsky, I.; Sideras, P.; Holland, J.; Davies, A.; Flinter, F.; Hammarstrom, L.; Kinnon, C.; Levinsky, R.; Bobrow, M.; et al., The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. *Nature* **1993**, 361, 226-233.

66. Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Eyck, L. F., Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *PNAS* **2006**, 103, 17783-17788.
67. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S., The protein kinase complement of the human genome. *Science* **2002**, 298, 1912-1934.
68. Heath, C. M.; Stahl, P. D.; Barbieri, M. A., Lipid kinases play crucial and multiple roles in membrane trafficking and signaling. *Histol Histopathol* **2003**, 18, 989-98.
69. Fabbro, D.; Cowan-Jacob, S. W.; Mobitz, H.; Martiny-Baron, G., Targeting cancer with small-molecular-weight kinase inhibitors. *Methods Mol Biol* **2012**, 795, 1-34.
70. Pearce, L. R.; Komander, D.; Alessi, D. R., The nuts and bolts of AGC protein kinases. *Nat Rev. Mol Cell Biol* **2010**, 11, 9-22.
71. Ma, W. W.; Adjei, A. A., Novel agents on the horizon for cancer therapy. *CA Cancer J Clin* **2009**, 59, 111-137.
72. Huang, M.; Shen, A.; Ding, J.; Geng, M., Molecularly targeted cancer therapy: some lessons from the past decade. *Trends Pharmacol Sci* **2014**, 35, 41-50.
73. Sun, C.; Bernards, R., Feedback and redundancy in receptor tyrosine kinase signaling: relevance to cancer therapies. *Trends Biochem Sci* **2014**, 39, 465-474.
74. Kikuchi, R.; Nakamura, K.; MacLauchlan, S.; Ngo, D. T.; Shimizu, I.; Fuster, J. J.; Katanasaka, Y.; Yoshida, S.; Qiu, Y.; Yamaguchi, T. P.; Matsushita, T.; Murohara, T.; Gokce, N.; Bates, D. O.; Hamburg, N. M.; Walsh, K., An antiangiogenic isoform of VEGF-A contributes to impaired vascularization in peripheral artery disease. *Nat Med* **2014**, 20, 1464-1471.
75. Muth, F.; Gunther, M.; Bauer, S. M.; Doring, E.; Fischer, S.; Maier, J.; Druckes, P.; Koppler, J.; Trappe, J.; Rothbauer, U.; Koch, P.; Laufer, S. A., Tetra-substituted pyridinylimidazoles as dual inhibitors of p38alpha mitogen-activated protein kinase and c-Jun N-terminal kinase 3 for potential treatment of neurodegenerative diseases. *J Med Chem* **2015**, 58, 443-456.
76. Clark, J. D.; Flanagan, M. E.; Telliez, J. B., Discovery and development of Janus kinase (JAK) inhibitors for inflammatory diseases. *J Med Chem* **2014**, 57, 5023-5038.

77. Barnes, P. J., New anti-inflammatory targets for chronic obstructive pulmonary disease. *Nature Rev Drug Discov* **2013**, 12, 543-559.
78. Banks, A. S.; McAllister, F. E.; Camporez, J. P.; Zushin, P. J.; Jureczak, M. J.; Laznik-Bogoslavski, D.; Shulman, G. I.; Gygi, S. P.; Spiegelman, B. M., An ERK/Cdk5 axis controls the diabetogenic actions of PPARgamma. *Nature* **2015**, 517, 391-395.
79. Futreal, P. A.; Coin, L.; Marshall, M.; Down, T.; Hubbard, T.; Wooster, R.; Rahman, N.; Stratton, M. R., A census of human cancer genes. *Nature Rev Cancer* **2004**, 4, 177-183.
80. Wu, P.; Nielsen, T. E.; Clausen, M. H., FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol Sci* **2015**, 36, 422-439.
81. Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M., Through the "gatekeeper door": exploiting the active kinase conformation. *J Med Chem* **2010**, 53, 2681-2694.
82. Liu Y, G. N. S., Rational design of inhibitors that bind to inactive kinase conformations. *Nat Chem Biol* **2006**, 2, 358-364.
83. Lamontanara, A. J.; Gencer, E. B.; Kuzyk, O.; Hantschel, O., Mechanisms of resistance to BCR-ABL and other kinase inhibitors. *Biochim Biophys Acta* **2013**, 1834, 1449-1459.
84. Ma, L.; Shan, Y.; Bai, R.; Xue, L.; Eide, C. A.; Ou, J.; Zhu, L. J.; Hutchinson, L.; Cerny, J.; Khoury, H. J.; Sheng, Z.; Druker, B. J.; Li, S.; Green, M. R., A therapeutically targetable mechanism of BCR-ABL-independent imatinib resistance in chronic myeloid leukemia. *Sci Transl Med* **2014**, 6, 252ra121.
85. Tan, Y.; Rouse, J.; Zhang, A.; Cariati, S.; Cohen, P.; Comb, M. J., FGF and stress regulate CREB and ATF-1 via a pathway involving p38 MAP kinase and MAPKAP kinase-2. *EMBO J* **1996**, 15, 4629-4642.
86. Heidenreich, O.; Neininger, A.; Schrott, G.; Zinck, R.; Cahill, M. A.; Engel, K.; Kotlyarov, A.; Kraft, R.; Kostka, S.; Gaestel, M.; Nordheim, A., MAPKAP kinase 2 phosphorylates serum response factor in vitro and in vivo. *J Biol Chem* **1999**, 274, 14434-14443.
87. Neufeld, B.; Grosse-Wilde, A.; Hoffmeyer, A.; Jordan, B. W.; Chen, P.; Dinev, D.; Ludwig, S.; Rapp, U. R., Serine/Threonine kinases 3pK and MAPK-activated protein kinase 2 interact with the basic helix-loop-helix transcription factor E47 and repress its transcriptional activity. *J Biol Chem* **2000**, 275, 20239-20242.

88. Stokoe, D.; Campbell, D. G.; Nakielny, S.; Hidaka, H.; Leever, S. J.; Marshall, C.; Cohen, P., MAPKAP kinase-2; a novel protein kinase activated by mitogen-activated protein kinase. *EMBO J* **1992**, 11, 3985-3994.
89. Sutherland, C.; Alterio, J.; Campbell, D. G.; Le Bourdelles, B.; Mallet, J.; Haavik, J.; Cohen, P., Phosphorylation and activation of human tyrosine hydroxylase in vitro by mitogen-activated protein (MAP) kinase and MAP-kinase-activated kinases 1 and 2. *Eur J Biochem* **1993**, 217, 715-722.
90. Huang, C. K.; Zhan, L.; Ai, Y.; Jongstra, J., LSP1 is the major substrate for mitogen-activated protein kinase-activated protein kinase 2 in human neutrophils. *J Biol Chem* **1997**, 272, 17-19.
91. Werz, O.; Klemm, J.; Samuelsson, B.; Radmark, O., 5-lipoxygenase is phosphorylated by p38 kinase-dependent MAPKAP kinases. *PNAS* **2000**, 97, 5261-5266.
92. van den Heuvel, S.; Harlow, E., Distinct roles for cyclin-dependent kinases in cell cycle control. *Science* **1993**, 262, 2050-2054.
93. Malumbres, M.; Barbacid, M., To cycle or not to cycle: a critical decision in cancer. *Nat Rev Cancer* **2001**, 1, 222-231.
94. Horiuchi, D.; Huskey, N. E.; Kusdra, L.; Wohlbold, L.; Merrick, K. A.; Zhang, C.; Creasman, K. J.; Shokat, K. M.; Fisher, R. P.; Goga, A., Chemical-genetic analysis of cyclin dependent kinase 2 function reveals an important role in cellular transformation by multiple oncogenic pathways. *PNAS* **2012**, 109, E1019-1027.
95. Ishizaki, T.; Maekawa, M.; Fujisawa, K.; Okawa, K.; Iwamatsu, A.; Fujita, A.; Watanabe, N.; Saito, Y.; Kakizuka, A.; Morii, N.; Narumiya, S., The small GTP-binding protein Rho binds to and activates a 160 kDa Ser/Thr protein kinase homologous to myotonic dystrophy kinase. *EMBO J* **1996**, 15, 1885-1893.
96. Amano, M.; Ito, M.; Kimura, K.; Fukata, Y.; Chihara, K.; Nakano, T.; Matsuura, Y.; Kaibuchi, K., Phosphorylation and activation of myosin by Rho-associated kinase (Rho-kinase). *J Biol Chem* **1996**, 271, 20246-20249.
97. Kawano, Y.; Fukata, Y.; Oshiro, N.; Amano, M.; Nakamura, T.; Ito, M.; Matsumura, F.; Inagaki, M.; Kaibuchi, K., Phosphorylation of myosin-binding subunit (MBS) of myosin phosphatase by Rho-kinase in vivo. *J Cell Biol* **1999**, 147, 1023-1038.
98. Kureishi, Y.; Kobayashi, S.; Amano, M.; Kimura, K.; Kanaide, H.; Nakano, T.; Kaibuchi, K.; Ito, M., Rho-associated kinase directly induces smooth muscle

- contraction through myosin light chain phosphorylation. *J Biol Chem* **1997**, 272, 12257-12260.
99. Ishizaki, T.; Naito, M.; Fujisawa, K.; Maekawa, M.; Watanabe, N.; Saito, Y.; Narumiya, S., p160ROCK, a Rho-associated coiled-coil forming protein kinase, works downstream of Rho and induces focal adhesions. *FEBS letters* **1997**, 404, 118-124.
 100. Hirose, M.; Ishizaki, T.; Watanabe, N.; Uehata, M.; Kranenburg, O.; Moolenaar, W. H.; Matsumura, F.; Maekawa, M.; Bito, H.; Narumiya, S., Molecular dissection of the Rho-associated protein kinase (p160ROCK)-regulated neurite remodeling in neuroblastoma N1E-115 cells. *J Cell Biol* **1998**, 141, 1625-1636.
 101. Yoshioka, K.; Matsumura, F.; Akedo, H.; Itoh, K., Small GTP-binding protein Rho stimulates the actomyosin system, leading to invasion of tumor cells. *J Biol Chem* **1998**, 273, 5146-5154.
 102. Khalil, R. A. Regulation of Vascular Smooth Muscle Function. In *Rho Kinase in Vascular Smooth Muscle*; Morgan & Claypool Life Sciences: CA, 2010.
 103. Croft, D. R.; Sahai, E.; Mavria, G.; Li, S.; Tsai, J.; Lee, W. M.; Marshall, C. J.; Olson, M. F., Conditional ROCK activation in vivo induces tumor cell dissemination and angiogenesis. *Cancer Res* **2004**, 64, 8994-9001.
 104. Mukai, Y.; Shimokawa, H.; Matoba, T.; Kandabashi, T.; Satoh, S.; Hiroki, J.; Kaibuchi, K.; Takeshita, A., Involvement of Rho-kinase in hypertensive vascular disease: a novel therapeutic target in hypertension. *FASEB J* **2001**, 15, 1062-1064.
 105. Shimokawa, H., Rho-kinase as a novel therapeutic target in treatment of cardiovascular diseases. *J Cardiovasc Pharmacol* **2002**, 39, 319-327.
 106. Ohki, S.; Iizuka, K.; Ishikawa, S.; Kano, M.; Dobashi, K.; Yoshii, A.; Shimizu, Y.; Mori, M.; Morishita, Y., A highly selective inhibitor of Rho-associated coiled-coil forming protein kinase, Y-27632, prolongs cardiac allograft survival of the BALB/c-to-C3H/He mouse model. *J Heart Lung Transplant* **2001**, 20, 956-963.
 107. Manning, B. D.; Cantley, L. C., AKT/PKB signaling: navigating downstream. *Cell* **2007**, 129, 1261-1274.
 108. Liu, P.; Cheng, H.; Roberts, T. M.; Zhao, J. J., Targeting the phosphoinositide 3-kinase pathway in cancer. *Nature Rev Drug Discov* **2009**, 8, 627-644.
 109. Addie, M.; Ballard, P.; Buttar, D.; Crafter, C.; Currie, G.; Davies, B. R.; Debreczeni, J.; Dry, H.; Dudley, P.; Greenwood, R.; Johnson, P. D.; Kettle, J. G.; Lane, C.; Lamont, G.; Leach, A.; Luke, R. W.; Morris, J.; Ogilvie, D.; Page, K.;

- Pass, M.; Pearson, S.; Ruston, L., Discovery of 4-amino-N-[(1S)-1-(4-chlorophenyl)-3-hydroxypropyl]-1-(7H-pyrrolo[2,3-d]pyrimidin-4-yl)piperidine-4-carboxamide (AZD5363), an orally bioavailable, potent inhibitor of Akt kinases. *J Med Chem* **2013**, 56, 2059-2073.
110. Baserga, R., The insulin-like growth factor-I receptor as a target for cancer therapy. *Expert Opin Ther Targets* **2005**, 9, 753-768.
111. Pollak, M. N.; Schernhammer, E. S.; Hankinson, S. E., Insulin-like growth factors and neoplasia. *Nat Rev Cancer* **2004**, 4, 505-518.
112. Yee, D., Targeting insulin-like growth factor pathways. *Br J Cancer* **2006**, 94, 465-468.
113. Hofmann, F.; Garcia-Echeverria, C., Blocking the insulin-like growth factor-I receptor as a strategy for targeting cancer. *Drug Discov Today* **2005**, 10, 1041-1047.
114. Garcia-Echeverria, C., Medicinal chemistry approaches to target the kinase activity of IGF-1R. *IDrugs* **2006**, 9, 415-419.
115. Bahr, C.; Groner, B., The insulin like growth factor-1 receptor (IGF-1R) as a drug target: novel approaches to cancer therapy. *Growth Horm IGF Res* **2004**, 14, 287-295.
116. Zhang, H.; Yee, D., The therapeutic potential of agents targeting the type I insulin-like growth factor receptor. *Expert Opin Investig Drugs* **2004**, 13, 1569-1577.
117. Himpe, E.; Kooijman, R., Insulin-like growth factor-I receptor signal transduction and the Janus Kinase/Signal Transducer and Activator of Transcription (JAK-STAT) pathway. *BioFactors* **2009**, 35, 76-81.
118. Kipreos, E. T.; Wang, J. Y., Differential phosphorylation of c-Abl in cell cycle determined by cdc2 kinase and phosphatase activity. *Science* **1990**, 248, 217-220.
119. Colicelli, J., ABL tyrosine kinases: evolution of function, regulation, and specificity. *Sci Signal* **2010**, 3, re6.
120. Becker, W. M., Kleinsmith, L.J., Hardin, J., Bertoni, G.P., *The World of the Cell*. San Francisco, CA, 2009.
121. Nowell, P. C. H., D.A, A minute chromosome in human chronic granulocytic leukemia. *Science* **1960**, 142, 1497.

122. Grossman, S. R.; Perez, M.; Kung, A. L.; Joseph, M.; Mansur, C.; Xiao, Z. X.; Kumar, S.; Howley, P. M.; Livingston, D. M., p300/MDM2 complexes participate in MDM2-mediated p53 degradation. *Mol Cell* **1998**, 2, 405-415.
123. Brooks, C. L.; Gu, W., p53 ubiquitination: Mdm2 and beyond. *Mol Cell* **2006**, 21, 307-315.
124. Fuchs, S. Y.; Adler, V.; Buschmann, T.; Wu, X.; Ronai, Z., Mdm2 association with p53 targets its ubiquitination. *Oncogene* **1998**, 17, 2543-2547.
125. Moll, U. M.; Petrenko, O., The MDM2-p53 interaction. *Mol Cancer Res* **2003**, 1, 1001-1008.
126. Momand, J.; Zambetti, G. P.; Olson, D. C.; George, D.; Levine, A. J., The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell* **1992**, 69, 1237-1245.
127. Marine, J. C.; Dyer, M. A.; Jochemsen, A. G., MDMX: from bench to bedside. *Journal Cell Sci* **2007**, 120, 371-378.
128. Gaglia, G.; Guan, Y.; Shah, J. V.; Lahav, G., Activation and control of p53 tetramerization in individual living cells. *PNAS* **2013**, 110, 15497-15501.
129. Bode, A. M.; Dong, Z., Post-translational modification of p53 in tumorigenesis. *Nat Rev Cancer* **2004**, 4, 793-805.
130. Sakaguchi, K.; Herrera, J. E.; Saito, S.; Miki, T.; Bustin, M.; Vassilev, A.; Anderson, C. W.; Appella, E., DNA damage activates p53 through a phosphorylation-acetylation cascade. *Genes Dev* **1998**, 12, 2831-2841.
131. Toledo, F.; Wahl, G. M., Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer* **2006**, 6, 909-923.
132. Brown, C. J.; Lain, S.; Verma, C. S.; Fersht, A. R.; Lane, D. P., Awakening guardian angels: drugging the p53 pathway. *Nat Rev Cancer* **2009**, 9, 862-873.
133. Follis, A. V.; Llambi, F.; Merritt, P.; Chipuk, J. E.; Green, D. R.; Kriwacki, R. W., Pin1-Induced Proline Isomerization in Cytosolic p53 Mediates BAX Activation and Apoptosis. *Mol Cell* **2015**, 59, 677-684.
134. Vousden, K. H.; Prives, C., Blinded by the Light: The Growing Complexity of p53. *Cell* **2009**, 137, 413-431.

135. Laptenko, O.; Tong, D. R.; Manfredi, J.; Prives, C., The Tail That Wags the Dog: How the Disordered C-Terminal Domain Controls the Transcriptional Activities of the p53 Tumor-Suppressor Protein. *Trends Biochem Sci* **2016**, 41, 1022-1034.
136. Laptenko, O.; Shiff, I.; Freed-Pastor, W.; Zupnick, A.; Mattia, M.; Freulich, E.; Shamir, I.; Kadouri, N.; Kahan, T.; Manfredi, J.; Simon, I.; Prives, C., The p53 C terminus controls site-specific DNA binding and promotes structural changes within the central DNA binding domain. *Mol Cell* **2015**, 57, 1034-1046.
137. Tidow, H.; Melero, R.; Mylonas, E.; Freund, S. M.; Grossmann, J. G.; Carazo, J. M.; Svergun, D. I.; Valle, M.; Fersht, A. R., Quaternary structures of tumor suppressor p53 and a specific p53 DNA complex. *PNAS* **2007**, 104, 12324-12329.
138. Bell, S.; Klein, C.; Muller, L.; Hansen, S.; Buchner, J., p53 contains large unstructured regions in its native state. *J Mol Biol* **2002**, 322, 917-27.
139. Gu, W.; Shi, X. L.; Roeder, R. G., Synergistic activation of transcription by CBP and p53. *Nature* **1997**, 387, 819-823.
140. Teufel, D. P.; Freund, S. M.; Bycroft, M.; Fersht, A. R., Four domains of p300 each bind tightly to a sequence spanning both transactivation subdomains of p53. *PNAS* **2007**, 104, 7009-7014.
141. Di Lello, P.; Jenkins, L. M.; Jones, T. N.; Nguyen, B. D.; Hara, T.; Yamaguchi, H.; Dikeakos, J. D.; Appella, E.; Legault, P.; Omichinski, J. G., Structure of the Tfb1/p53 complex: Insights into the interaction between the p62/Tfb1 subunit of TFIIF and the activation domain of p53. *Mol Cell* **2006**, 22, 731-740.
142. Dawson, R.; Muller, L.; Dehner, A.; Klein, C.; Kessler, H.; Buchner, J., The N-terminal domain of p53 is natively unfolded. *J Mol Biol* **2003**, 332, 1131-1141.
143. Lee, H.; Mok, K. H.; Muhandiram, R.; Park, K. H.; Suk, J. E.; Kim, D. H.; Chang, J.; Sung, Y. C.; Choi, K. Y.; Han, K. H., Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J Biol Chem* **2000**, 275, 29426-29432.
144. Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. P., Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **1996**, 274, 948-953.
145. Schon, O.; Friedler, A.; Bycroft, M.; Freund, S. M.; Fersht, A. R., Molecular mechanism of the interaction between MDM2 and p53. *J Mol Biol* **2002**, 323, 491-501.
146. Sugase, K.; Dyson, H. J.; Wright, P. E., Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **2007**, 447, 1021-1025.

147. Joerger, A. C.; Fersht, A. R., Structural biology of the tumor suppressor p53. *Annu Rev Biochem* **2008**, 77, 557-582.
148. Ferreon, J. C.; Lee, C. W.; Arai, M.; Martinez-Yamout, M. A.; Dyson, H. J.; Wright, P. E., Cooperative regulation of p53 by modulation of ternary complex formation with CBP/p300 and HDM2. *PNAS* **2009**, 106, 6591-6596.
149. Toledo, F.; Krummel, K. A.; Lee, C. J.; Liu, C. W.; Rodewald, L. W.; Tang, M.; Wahl, G. M., A mouse p53 mutant lacking the proline-rich domain rescues Mdm4 deficiency and provides insight into the Mdm2-Mdm4-p53 regulatory network. *Cancer Cell* **2006**, 9, 273-285.
150. Natan, E.; Baloglu, C.; Pagel, K.; Freund, S. M.; Morgner, N.; Robinson, C. V.; Fersht, A. R.; Joerger, A. C., Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. *J Mol Biol* **2011**, 409, 358-368.
151. Cho, Y.; Gorina, S.; Jeffrey, P. D.; Pavletich, N. P., Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* **1994**, 265, 346-355.
152. Petty, T. J.; Emamzadah, S.; Costantino, L.; Petkova, I.; Stavridi, E. S.; Saven, J. G.; Vauthey, E.; Halazonetis, T. D., An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. *EMBO J* **2011**, 30, 2167-2176.
153. Canadillas, J. M.; Tidow, H.; Freund, S. M.; Rutherford, T. J.; Ang, H. C.; Fersht, A. R., Solution structure of p53 core domain: structural basis for its instability. *PNAS* **2006**, 103, 2109-2114.
154. Wang, Y.; Schwedes, J. F.; Parks, D.; Mann, K.; Tegtmeyer, P., Interaction of p53 with its consensus DNA-binding site. *Mol Cell Biol* **1995**, 15, 2157-2165.
155. Weinberg, R. L.; Veprintsev, D. B.; Fersht, A. R., Cooperative binding of tetrameric p53 to DNA. *J Mol Biol* **2004**, 341, 1145-1159.
156. Emamzadah, S.; Tropia, L.; Halazonetis, T. D., Crystal structure of a multidomain human p53 tetramer bound to the natural CDKN1A (p21) p53-response element. *Mol Cancer Res* **2011**, 9, 1493-1499.
157. Kitayner, M.; Rozenberg, H.; Kessler, N.; Rabinovich, D.; Shaulov, L.; Haran, T. E.; Shakked, Z., Structural basis of DNA recognition by p53 tetramers. *Mol Cell* **2006**, 22, 741-753.

158. Arbely, E.; Natan, E.; Brandt, T.; Allen, M. D.; Veprintsev, D. B.; Robinson, C. V.; Chin, J. W.; Joerger, A. C.; Fersht, A. R., Acetylation of lysine 120 of p53 endows DNA-binding specificity at effective physiological salt concentration. *PNAS* **2011**, 108, 8251-8256.
159. Melero, R.; Rajagopalan, S.; Lazaro, M.; Joerger, A. C.; Brandt, T.; Veprintsev, D. B.; Lasso, G.; Gil, D.; Scheres, S. H.; Carazo, J. M.; Fersht, A. R.; Valle, M., Electron microscopy studies on the quaternary structure of p53 reveal different binding modes for p53 tetramers in complex with DNA. *PNAS* **2011**, 108, 557-562.
160. Weinberg, R. L.; Veprintsev, D. B.; Bycroft, M.; Fersht, A. R., Comparative binding of p53 to its promoter and DNA recognition elements. *J Mol Biol* **2005**, 348, 589-596.
161. Soussi, T.; Asselain, B.; Hamroun, D.; Kato, S.; Ishioka, C.; Claustres, M.; Beroud, C., Meta-analysis of the p53 mutation database for mutant p53 biological activity reveals a methodologic bias in mutation detection. *Clin Cancer Res* **2006**, 12, 62-69.
162. Shaulsky, G.; Goldfinger, N.; Ben-Ze'ev, A.; Rotter, V., Nuclear accumulation of p53 protein is mediated by several nuclear localization signals and plays a role in tumorigenesis. *Mol Cell Biol* **1990**, 10, 6565-6577.
163. Liang, S. H.; Clarke, M. F., A bipartite nuclear localization signal is required for p53 nuclear import regulated by a carboxyl-terminal domain. *J Biol Chem* **1999**, 274, 32699-32703.
164. Marchenko, N. D.; Hanel, W.; Li, D.; Becker, K.; Reich, N.; Moll, U. M., Stress-mediated nuclear stabilization of p53 is regulated by ubiquitination and importin- α 3 binding. *Cell Death Differ* **2010**, 17, 255-267.
165. Marine, J. C., p53 stabilization: the importance of nuclear import. *Cell Death Differ* **2010**, 17, 191-192.
166. Clore, G. M.; Ernst, J.; Clubb, R.; Omichinski, J. G.; Kennedy, W. M.; Sakaguchi, K.; Appella, E.; Gronenborn, A. M., Refined solution structure of the oligomerization domain of the tumour suppressor p53. *Nat Struct Biol* **1995**, 2, 321-333.
167. Davison, T. S.; Nie, X.; Ma, W.; Lin, Y.; Kay, C.; Benchimol, S.; Arrowsmith, C. H., Structure and functionality of a designed p53 dimer. *J Mol Biol* **2001**, 307, 605-617.

168. Jeffrey, P. D.; Gorina, S.; Pavletich, N. P., Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms. *Science* **1995**, 267, 1498-1502.
169. Mittl, P. R.; Chene, P.; Grutter, M. G., Crystallization and structure solution of p53 (residues 326-356) by molecular replacement using an NMR model as template. *Acta Cryst D* **1998**, 54, 86-89.
170. Sakamoto, H.; Lewis, M. S.; Kodama, H.; Appella, E.; Sakaguchi, K., Specific sequences from the carboxyl terminus of human p53 gene product form anti-parallel tetramers in solution. *PNAS* **1994**, 91, 8974-8978.
171. Sturzbecher, H. W.; Brain, R.; Addison, C.; Rudge, K.; Remm, M.; Grimaldi, M.; Keenan, E.; Jenkins, J. R., A C-terminal alpha-helix plus basic region motif is the major structural determinant of p53 tetramerization. *Oncogene* **1992**, 7, 1513-1523.
172. Rajagopalan, S.; Huang, F.; Fersht, A. R., Single-Molecule characterization of oligomerization kinetics and equilibria of the tumor suppressor p53. *Nucleic Acids Res* **2011**, 39, 2294-2303.
173. Pant, V.; Lozano, G., Limiting the power of p53 through the ubiquitin proteasome pathway. *Genes Dev* **2014**, 28, 1739-1751.
174. Demir, O.; Jeong, P. U.; Amaro, R. E., Full-length p53 tetramer bound to DNA and its quaternary dynamics. *Oncogene* **2017**, 36, 1451-1460.
175. Friedler, A.; Veprintsev, D. B.; Freund, S. M.; von Glos, K. I.; Fersht, A. R., Modulation of binding of DNA to the C-terminal domain of p53 by acetylation. *Structure* **2005**, 13, 629-636.
176. Ahn, J.; Prives, C., The C-terminus of p53: the more you learn the less you know. *Nat Struct Biol* **2001**, 8, 730-732.
177. Reed, S. M.; Quelle, D. E., p53 Acetylation: Regulation and Consequences. *Cancers* **2014**, 7, 30-69.
178. Hupp, T. R.; Meek, D. W.; Midgley, C. A.; Lane, D. P., Regulation of the specific DNA binding function of p53. *Cell* **1992**, 71, 875-886.
179. Ayed, A.; Mulder, F. A.; Yi, G. S.; Lu, Y.; Kay, L. E.; Arrowsmith, C. H., Latent and active p53 are identical in conformation. *Nature structural biology* **2001**, 8, 756-60.
180. Jayaraman, L.; Prives, C., Covalent and noncovalent modifiers of the p53 protein. *Cell Mol Life Sci* **1999**, 55, 76-87.

181. McKinney, K.; Mattia, M.; Gottifredi, V.; Prives, C., p53 linear diffusion along DNA requires its C terminus. *Mol Cell* **2004**, 16, 413-424.
182. Espinosa, J. M.; Emerson, B. M., Transcriptional regulation by p53 through intrinsic DNA/chromatin binding and site-directed cofactor recruitment. *Mol Cell* **2001**, 8, 57-69.
183. Gohler, T.; Reimann, M.; Cherny, D.; Walter, K.; Warnecke, G.; Kim, E.; Deppert, W., Specific interaction of p53 with target binding sites is determined by DNA conformation and is regulated by the C-terminal domain. *J Biol Chem* **2002**, 277, 41192-41203.
184. Kim, H.; Kim, K.; Choi, J.; Heo, K.; Baek, H. J.; Roeder, R. G.; An, W., p53 requires an intact C-terminal domain for DNA binding and transactivation. *J Mol Biol* **2012**, 415, 843-854.
185. McKinney, K.; Prives, C., Efficient specific DNA binding by p53 requires both its central and C-terminal domains as revealed by studies with high-mobility group 1 protein. *Mol Cell Biol* **2002**, 22, 6797-6808.
186. Palecek, E.; Brazda, V.; Jagelska, E.; Pecinka, P.; Karlovska, L.; Brazdova, M., Enhancement of p53 sequence-specific binding by DNA supercoiling. *Oncogene* **2004**, 23, 2119-2127.
187. Tafvizi, A.; Huang, F.; Fersht, A. R.; Mirny, L. A.; van Oijen, A. M., A single-molecule characterization of p53 search on DNA. *PNAS* **2011**, 108, 563-568.
188. Terakawa, T.; Kenzaki, H.; Takada, S., p53 searches on DNA by rotation-uncoupled sliding at C-terminal tails and restricted hopping of core domains. *JACS* **2012**, 134, 14555-14562.
189. Harris, C. C.; Hollstein, M., Clinical implications of the p53 tumor-suppressor gene. *New Engl J Med* **1993**, 329, 1318-1327.
190. Petitjean, A.; Achatz, M. I.; Borresen-Dale, A. L.; Hainaut, P.; Olivier, M., TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene* **2007**, 26, 2157-2165.
191. Strano, S.; Dell'Orso, S.; Di Agostino, S.; Fontemaggi, G.; Sacchi, A.; Blandino, G., Mutant p53: an oncogenic transcription factor. *Oncogene* **2007**, 26, 2212-2219.
192. Sigal, A.; Rotter, V., Oncogenic mutations of the p53 tumor suppressor: the demons of the guardian of the genome. *Cancer Res* **2000**, 60, 6788-6793.

193. Hinds, P. W.; Finlay, C. A.; Quartin, R. S.; Baker, S. J.; Fearon, E. R.; Vogelstein, B.; Levine, A. J., Mutant p53 DNA clones from human colon carcinomas cooperate with ras in transforming primary rat cells: a comparison of the "hot spot" mutant phenotypes. *Cell Growth Differ* **1990**, 1, 571-580.
194. Stephen, C. W.; Lane, D. P., Mutant Conformation of P53 - Precise Epitope Mapping Using a Filamentous Phage Epitope Library. *J Mol Biol* **1992**, 225, 577-583.
195. Pavletich, N. P.; Chambers, K. A.; Pabo, C. O., The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Genes Dev* **1993**, 7, 2556-2564.
196. Bargonetti, J.; Manfredi, J. J.; Chen, X.; Marshak, D. R.; Prives, C., A proteolytic fragment from the central region of p53 has marked sequence-specific DNA-binding activity when generated from wild-type but not from oncogenic mutant p53 protein. *Genes Dev* **1993**, 7, 2565-2574.
197. Raycroft, L.; Schmidt, J. R.; Yoas, K.; Hao, M. M.; Lozano, G., Analysis of p53 mutants for transcriptional activity. *Mol Cell Biol* **1991**, 11, 6067-6074.
198. Brosh, R.; Rotter, V., When mutants gain new powers: news from the mutant p53 field. *Nat Rev Cancer* **2009**, 9, 701-713.
199. Di Como, C. J.; Gaiddon, C.; Prives, C., p73 function is inhibited by tumor-derived p53 mutants in mammalian cells. *Mol Cell Biol* **1999**, 19, 1438-1449.
200. Gaiddon, C.; Lokshin, M.; Ahn, J.; Zhang, T.; Prives, C., A subset of tumor-derived mutant forms of p53 down-regulate p63 and p73 through a direct interaction with the p53 core domain. *Mol Cell Biol* **2001**, 21, 1874-1887.
201. Irwin, M. S.; Kondo, K.; Marin, M. C.; Cheng, L. S.; Hahn, W. C.; Kaelin, W. G., Jr., Chemosensitivity linked to p73 function. *Cancer Cell* **2003**, 3, 403-410.
202. Marin, M. C.; Jost, C. A.; Brooks, L. A.; Irwin, M. S.; O'Nions, J.; Tidy, J. A.; James, N.; McGregor, J. M.; Harwood, C. A.; Yulug, I. G.; Vousden, K. H.; Allday, M. J.; Gusterson, B.; Ikawa, S.; Hinds, P. W.; Crook, T.; Kaelin, W. G., Jr., A common polymorphism acts as an intragenic modifier of mutant p53 behaviour. *Nat Genet* **2000**, 25, 47-54.
203. Strano, S.; Munarriz, E.; Rossi, M.; Castagnoli, L.; Shaul, Y.; Sacchi, A.; Oren, M.; Sudol, M.; Cesareni, G.; Blandino, G., Physical interaction with Yes-associated protein enhances p73 transcriptional activity. *J Biol Chem* **2001**, 276, 15164-15173.

204. Sampath, J.; Sun, D.; Kidd, V. J.; Grenet, J.; Gandhi, A.; Shapiro, L. H.; Wang, Q.; Zambetti, G. P.; Schuetz, J. D., Mutant p53 cooperates with ETS and selectively up-regulates human MDR1 not MRP1. *J Biol Chem* **2001**, 276, 39359-39367.
205. Adorno, M.; Cordenonsi, M.; Montagner, M.; Dupont, S.; Wong, C.; Hann, B.; Solari, A.; Bobisse, S.; Rondina, M. B.; Guzzardo, V.; Parenti, A. R.; Rosato, A.; Bicciato, S.; Balmain, A.; Piccolo, S., A Mutant-p53/Smad complex opposes p63 to empower TGFbeta-induced metastasis. *Cell* **2009**, 137, 87-98.
206. Muller, P. A.; Caswell, P. T.; Doyle, B.; Iwanicki, M. P.; Tan, E. H.; Karim, S.; Lukashchuk, N.; Gillespie, D. A.; Ludwig, R. L.; Gosselin, P.; Cromer, A.; Brugge, J. S.; Sansom, O. J.; Norman, J. C.; Vousden, K. H., Mutant p53 drives invasion by promoting integrin recycling. *Cell* **2009**, 139, 1327-1341.
207. Davison, T. S.; Vagner, C.; Kaghad, M.; Ayed, A.; Caput, D.; Arrowsmith, C. H., p73 and p63 are homotetramers capable of weak heterotypic interactions with each other but not with p53. *J Biol Chem* **1999**, 274, 18709-18714.
208. Weisz, L.; Oren, M.; Rotter, V., Transcription regulation by mutant p53. *Oncogene* **2007**, 26, 2202-2211.
209. Chin, K. V.; Ueda, K.; Pastan, I.; Gottesman, M. M., Modulation of activity of the promoter of the human MDR1 gene by Ras and p53. *Science* **1992**, 255, 459-462.
210. Lin, J.; Teresky, A. K.; Levine, A. J., Two critical hydrophobic amino acids in the N-terminal domain of the p53 protein are required for the gain of function phenotypes of human p53 mutants. *Oncogene* **1995**, 10, 2387-2390.
211. Strauss, B. E.; Haas, M., The region 3' to the major transcriptional start site of the MDR1 downstream promoter mediates activation by a subset of mutant P53 proteins. *Biochem Biophys Res Commun* **1995**, 217, 333-340.
212. Lee, Y. I.; Lee, S.; Das, G. C.; Park, U. S.; Park, S. M.; Lee, Y. I., Activation of the insulin-like growth factor II transcription by aflatoxin B1 induced p53 mutant 249 is caused by activation of transcription complexes; implications for a gain-of-function during the formation of hepatocellular carcinoma. *Oncogene* **2000**, 19, 3717-3726.
213. Pugacheva, E. N.; Ivanov, A. V.; Kravchenko, J. E.; Kopnin, B. P.; Levine, A. J.; Chumakov, P. M., Novel gain of function activity of p53 mutants: activation of the dUTPase gene expression leading to resistance to 5-fluorouracil. *Oncogene* **2002**, 21, 4595-4600.

214. Bossi, G.; Marampon, F.; Maor-Aloni, R.; Zani, B.; Rotter, V.; Oren, M.; Strano, S.; Blandino, G.; Sacchi, A., Conditional RNA interference in vivo to study mutant p53 oncogenic gain of function on tumor malignancy. *Cell Cycle* **2008**, *7*, 1870-1879.
215. Sankala, H.; Vaughan, C.; Wang, J.; Deb, S.; Graves, P. R., Upregulation of the mitochondrial transport protein, Tim50, by mutant p53 contributes to cell growth and chemoresistance. *Arch Biochem Biophys* **2011**, *512*, 52-60.
216. Lavra, L.; Ulivieri, A.; Rinaldo, C.; Dominici, R.; Volante, M.; Luciani, E.; Bartolazzi, A.; Frasca, F.; Soddu, S.; Sciacchitano, S., Gal-3 is stimulated by gain-of-function p53 mutations and modulates chemoresistance in anaplastic thyroid carcinomas. *J Pathol* **2009**, *218*, 66-75.
217. Scian, M. J.; Stagliano, K. E.; Anderson, M. A.; Hassan, S.; Bowman, M.; Miles, M. F.; Deb, S. P.; Deb, S., Tumor-derived p53 mutants induce NF-kappaB2 gene expression. *Mol Cell Biol* **2005**, *25*, 10097-10110.
218. Vaughan, C. A.; Singh, S.; Windle, B.; Sankala, H. M.; Graves, P. R.; Andrew Yeudall, W.; Deb, S. P.; Deb, S., p53 mutants induce transcription of NF-kappaB2 in H1299 cells through CBP and STAT binding on the NF-kappaB2 promoter and gain of function activity. *Arch Biochem Biophys* **2012**, *518*, 79-88.
219. Christophorou, M. A.; Martin-Zanca, D.; Soucek, L.; Lawlor, E. R.; Brown-Swigart, L.; Verschuren, E. W.; Evan, G. I., Temporal dissection of p53 function in vitro and in vivo. *Nat Genet* **2005**, *37*, 718-726.
220. Christophorou, M. A.; Ringshausen, I.; Finch, A. J.; Swigart, L. B.; Evan, G. I., The pathological response to DNA damage does not contribute to p53-mediated tumour suppression. *Nature* **2006**, *443*, 214-217.
221. Xue, W.; Zender, L.; Miething, C.; Dickins, R. A.; Hernando, E.; Krizhanovsky, V.; Cordon-Cardo, C.; Lowe, S. W., Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature* **2007**, *445*, 656-660.
222. Kenzelmann Broz, D.; Attardi, L. D., In vivo analysis of p53 tumor suppressor function using genetically engineered mouse models. *Carcinogenesis* **2010**, *31*, 1311-1318.
223. Wang, Y.; Suh, Y. A.; Fuller, M. Y.; Jackson, J. G.; Xiong, S.; Terzian, T.; Quintas-Cardama, A.; Bankson, J. A.; El-Naggar, A. K.; Lozano, G., Restoring expression of wild-type p53 suppresses tumor growth but does not cause tumor regression in mice with a p53 missense mutation. *J Clin Invest* **2011**, *121*, 893-904.

224. Chen, F.; Wang, W.; El-Deiry, W. S., Current strategies to target p53 in cancer. *Biochem Pharmacol* **2010**, 80, 724-730.
225. Brown, C. J.; Cheok, C. F.; Verma, C. S.; Lane, D. P., Reactivation of p53: from peptides to small molecules. *Trends in Pharmacol Sci* **2011**, 32, 53-62.
226. Di, J.; Zhang, Y.; Zheng, J., Reactivation of p53 by inhibiting Mdm2 E3 ligase: a novel antitumor approach. *Curr Cancer Drug Targets* **2011**, 11, 987-994.
227. Vu, B. T.; Vassilev, L., Small-molecule inhibitors of the p53-MDM2 interaction. *Curr Top Microbiol Immunol* **2011**, 348, 151-172.
228. Foster, B. A.; Coffey, H. A.; Morin, M. J.; Rastinejad, F., Pharmacological rescue of mutant p53 conformation and function. *Science* **1999**, 286, 2507-2510.
229. Rippin, T. M.; Bykov, V. J.; Freund, S. M.; Selivanova, G.; Wiman, K. G.; Fersht, A. R., Characterization of the p53-rescue drug CP-31398 in vitro and in living cells. *Oncogene* **2002**, 21, 2119-2129.
230. Bykov, V. J.; Issaeva, N.; Shilov, A.; Hultcrantz, M.; Pugacheva, E.; Chumakov, P.; Bergman, J.; Wiman, K. G.; Selivanova, G., Restoration of the tumor suppressor function to mutant p53 by a low-molecular-weight compound. *Nat Med* **2002**, 8, 282-288.
231. Wiman, K. G., Pharmacological reactivation of mutant p53: from protein structure to the cancer patient. *Oncogene* **2010**, 29, 4245-4252.
232. National Cancer Institute., Clinical Trials Using PRIMA-1 Analog APR-246, <http://www.cancer.gov/about-cancer/treatment/clinical-trials/intervention/C85465>.
233. Kravchenko, J. E.; Ilyinskaya, G. V.; Komarov, P. G.; Agapova, L. S.; Kochetkov, D. V.; Strom, E.; Frolova, E. I.; Kovriga, I.; Gudkov, A. V.; Feinstein, E.; Chumakov, P. M., Small-molecule RETRA suppresses mutant p53-bearing cancer cells through a p73-dependent salvage pathway. *PNAS* **2008**, 105, 6302-6307.
234. Li, D.; Marchenko, N. D.; Moll, U. M., SAHA shows preferential cytotoxicity in mutant p53 cancer cells by destabilizing mutant p53 through inhibition of the HDAC6-Hsp90 chaperone axis. *Cell Death Differ* **2011**, 18, 1904-1913.
235. Li, D.; Marchenko, N. D.; Schulz, R.; Fischer, V.; Velasco-Hernandez, T.; Talos, F.; Moll, U. M., Functional inactivation of endogenous MDM2 and CHIP by HSP90 causes aberrant stabilization of mutant p53 in human cancer cells. *Mol Cancer Res* **2011**, 9, 577-588.

236. Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R., Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. *PNAS* **2008**, 105, 10360-10365.
237. Wassman, C. D.; Baronio, R.; Demir, O.; Wallentine, B. D.; Chen, C. K.; Hall, L. V.; Salehi, F.; Lin, D. W.; Chung, B. P.; Hatfield, G. W.; Richard Chamberlin, A.; Luecke, H.; Lathrop, R. H.; Kaiser, P.; Amaro, R. E., Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nature Commun* **2013**, 4, 1407.
238. Vendruscolo, M., Determination of conformationally heterogeneous states of proteins. *Curr Opin Struct Biol* **2007**, 17, 15-20.
239. Boehr, D. D.; Nussinov, R.; Wright, P. E., The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* **2009**, 5, 789-796.
240. Osawa, M.; Takeuchi, K.; Ueda, T.; Nishida, N.; Shimada, I., Functional dynamics of proteins revealed by solution NMR. *Curr Opin Struct Biol* **2012**, 22, 660-669.
241. Zhuravlev, P. I.; Papoian, G. A., Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. *Q Rev Biophys* **2010**, 43, 295-332.
242. Weber, G., Energetics of ligand binding to proteins. *Adv Protein Chem* **1975**, 29, 1-83.
243. Göbl, C.; Tjandra, N., Application of Solution NMR Spectroscopy to Study Protein Dynamics. *Entropy* **2012**, 14, 581-598.
244. Perutz, M. F.; Mathews, F. S., An x-ray study of azide methaemoglobin. *J Mol Biol* **1966**, 21, 199-202.
245. Joseph, D.; Petsko, G. A.; Karplus, M., Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop. *Science* **1990**, 249, 1425-1428.
246. Faber, H. R.; Matthews, B. W., A mutant T4 lysozyme displays five different crystal conformations. *Nature* **1990**, 348, 263-266.
247. Bennett, W. S., Jr.; Steitz, T. A., Glucose-induced conformational change in yeast hexokinase. *PNAS* **1978**, 75, 4848-4852.

248. Popovych, N.; Sun, S.; Ebright, R. H.; Kalodimos, C. G., Dynamically driven protein allostery. *Nat Struct Mol Biol* **2006**, 13, 831-838.
249. Manley, G.; Loria, J. P., NMR insights into protein allostery. *Arch Biochem Biophys* **2012**, 519, 223-231.
250. Motlagh, H. N.; Li, J.; Thompson, E. B.; Hilser, V. J., Interplay between allostery and intrinsic disorder in an ensemble. *Biochem Soc Trans* **2012**, 40, 975-980.
251. Tsai, C. J.; Nussinov, R., A unified view of "how allostery works". *PLOS Comput Biol* **2014**, 10, e1003394.
252. Corey, R. B., Pauling, L., Molecular models of amino acids, peptides, and proteins. *Rev Sci Instrum* **1953**, 24, 621-627.
253. McCammon, J. A., Harvey, S.C., *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press: Cambridge, 1987.
254. Cornell, W. D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., Kollman, P.A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *JACS* **1995**, 117, 5179-5197.
255. Durrant, J. D.; McCammon, J. A., Molecular dynamics simulations and drug discovery. *BMC Biol* **2011**, 9, 71.
256. Lennard-Jones, J. E., On the determination of molecular fields. - II. From the equation of state of a gas. *Proc R Soc Lond A* **1924**, 106, 463-477.
257. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004**, 25, 1157-1174.
258. Brooks, B. R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M., CHARMM - a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* **1983**, 4, 187-217.
259. Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F., The GROMOS software for biomolecular simulation: GROMOS05. *J Comput Chem* **2005**, 26, 1719-1751.
260. Martin-Garcia, F.; Papaleo, E.; Gomez-Puertas, P.; Boomsma, W.; Lindorff-Larsen, K., Comparing molecular dynamics force fields in the essential subspace. *PloS One* **2015**, 10, e0121114.

261. Adcock, S. A.; McCammon, J. A., Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev* **2006**, 106, 1589-1615.
262. Duan, Y.; Kollman, P. A., Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, 282, 740-744.
263. Swendsen, R. H.; Wang, J. S., Replica Monte Carlo simulation of spin glasses. *Physical Rev Lett* **1986**, 57, 2607-2609.
264. Sugita, Y. a. O., Y., Replica exchange molecular dynamics method for protein folding simulation. *Chem Phys Lett* **1999**, 314, 141-151.
265. Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H., Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *JACS* **2007**, 129, 1179-1189.
266. Best, R. B.; Buchete, N. V.; Hummer, G., Are current molecular dynamics force fields too helical? *Biophys J* **2008**, 95, L07-9.
267. Jiang, F.; Han, W.; Wu, Y. D., Influence of side chain conformations on local conformational features of amino acids and implication for force field development. *J Phys Chem B* **2010**, 114, 5840-5850.
268. Jiang, F.; Han, W.; Wu, Y. D., The intrinsic conformational features of amino acids from a protein coil library and their applications in force field development. *Phys Chem Chem Phys* **2013**, 15, 3413-3428.
269. Cino, E. A.; Choy, W. Y.; Karttunen, M., Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J Chem Theory Comput* **2012**, 8, 2725-2740.
270. Gnanakaran, S.; Garcia, A. E., Helix-coil transition of alanine peptides in water: force field dependence on the folded and unfolded structures. *Proteins* **2005**, 59, 773-782.
271. Best, R. B.; Hummer, G., Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B* **2009**, 113, 9004-9015.
272. Kaminski, G. A., Friesner, R.A., Tirado-Rives, J., Jorgensen, W.L., Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* **2001**, 2, 6474-6487.

273. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, 65, 712-725.
274. MacKerell, A. D., Jr.; Banavali, N.; Foloppe, N., Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, 56, 257-265.
275. MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., 3rd, Improved treatment of the protein backbone in empirical force fields. *JACS* **2004**, 126, 698-699.
276. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, 78, 1950-1958.
277. Beauchamp, K. A.; Lin, Y. S.; Das, R.; Pande, V. S., Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J Chem Theory Comput* **2012**, 8, 1409-1414.
278. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; Mackerell, A. D., Jr., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput* **2012**, 8, 3257-3273.
279. Vitalini, F.; Mey, A. S.; Noe, F.; Keller, B. G., Dynamic properties of force fields. *J Chem Phys* **2015**, 142, 084101.
280. Lange, O. F.; van der Spoel, D.; de Groot, B. L., Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. *Biophys J* **2010**, 99, 647-655.
281. Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E., Systematic validation of protein force fields against experimental data. *PloS One* **2012**, 7, e32131.
282. Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A., Discovery of a novel binding trench in HIV integrase. *J Med Chem* **2004**, 47, 1879-1881.
283. Hazuda, D. J.; Anthony, N. J.; Gomez, R. P.; Jolly, S. M.; Wai, J. S.; Zhuang, L.; Fisher, T. E.; Embrey, M.; Guare, J. P., Jr.; Egbertson, M. S.; Vacca, J. P.; Huff, J. R.; Felock, P. J.; Witmer, M. V.; Stillmock, K. A.; Danovich, R.; Grobler, J.; Miller, M. D.; Espeseth, A. S.; Jin, L.; Chen, I. W.; Lin, J. H.; Kassahun, K.; Ellis, J. D.; Wong, B. K.; Xu, W.; Pearson, P. G.; Schleif, W. A.; Cortese, R.; Emini, E.; Summa, V.; Holloway, M. K.; Young, S. D., A naphthyridine carboxamide

- provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. *PNAS* **2004**, 101, 11233-11238.
284. Savarino, A., A historical sketch of the discovery and development of HIV-1 integrase inhibitors. *Expert Opin Investig Drugs* **2006**, 15, 1507-1522.
285. Actavalon; <http://www.actavalon.com/v2/index.php>.
286. Amaro, R. E.; Baron, R.; McCammon, J. A., An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J Comput Aided Mol Des* **2008**, 22, 693-705.
287. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A., The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers* **2003**, 68, 47-62.
288. Swift, R. V.; Jusoh, S. A.; Offutt, T. L.; Li, E. S.; Amaro, R. E., Knowledge-Based Methods To Train and Optimize Virtual Screening Ensembles. *J Chem Inf Model* **2016**, 56, 830-842.
289. Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A., Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *JACS* **2002**, 124, 5632-5633.
290. Amaro, R. E.; Schnauffer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A., Discovery of drug-like inhibitors of an essential RNA-editing ligase in *Trypanosoma brucei*. *PNAS* **2008**, 105, 17278-17283.
291. Durrant, J. D.; Hall, L.; Swift, R. V.; Landon, M.; Schnauffer, A.; Amaro, R. E., Novel naphthalene-based inhibitors of *Trypanosoma brucei* RNA editing ligase 1. *PLoS Negl Trop Dis* **2010**, 4, e803.
292. Durrant, J. D.; Urbaniak, M. D.; Ferguson, M. A.; McCammon, J. A., Computer-aided identification of *Trypanosoma brucei* uridine diphosphate galactose 4'-epimerase inhibitors: toward the development of novel therapies for African sleeping sickness. *J Med Chem* **2010**, 53, 5025-5032.
293. Durrant, J. D., Cao, R., Gorfe, A.A., Zhu, W., Li, J., Sankovsky, A., Oldfield, E., McCammon, J.A., Non-bisphosphonate inhibitors of isoprenoid biosynthesis identified via computer-aided drug design. *Chem Biol Drug Des* **2011**, 78, 323-332.
294. Wang, Y.; Hess, T. N.; Jones, V.; Zhou, J. Z.; McNeil, M. R.; Andrew McCammon, J., Novel inhibitors of *Mycobacterium tuberculosis* dTDP-6-deoxy-

- L-lyxo-4-hexulose reductase (RmlD) identified by virtual screening. *Bioorganic Med Chem Lett* **2011**, 21, 7064-7067.
295. Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S., Alchemical free energy methods for drug discovery: progress and challenges. *Curr Opin Struct Biol* **2011**, 21, 150-160.
296. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Rev Drug Discov* **2004**, 3, 935-949.
297. Schwab, F.; van Gunsteren, W. F.; Zagrovic, B., Computational study of the mechanism and the relative free energies of binding of anticholesteremic inhibitors to squalene-hopene cyclase. *Biochemistry* **2008**, 47, 2945-2951.
298. Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domaoal, R. A.; Wang, L.; Anderson, K. S.; Jorgensen, W. L., FEP-guided selection of bicyclic heterocycles in lead optimization for non-nucleoside inhibitors of HIV-1 reverse transcriptase. *JACS* **2006**, 128, 15372-15373.
299. Tembe, B. L., McCammon, J.A., Ligand receptor interactions. *Comput Chem* **1984**, 8, 281-283.
300. Chipot, C.; Rozanska, X.; Dixit, S. B., Can free energy calculations be fast and accurate at the same time? Binding of low-affinity, non-peptide inhibitors to the SH2 domain of the src protein. *J Comput Aided Mol Des* **2005**, 19, 765-770.
301. Deng, Y.; Roux, B., Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J Chem Theory Comput* **2006**, 2, 1255-1273.
302. Wang, J.; Deng, Y.; Roux, B., Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys J* **2006**, 91, 2798-2814.
303. Zeevaart, J. G.; Wang, L.; Thakur, V. V.; Leung, C. S.; Tirado-Rives, J.; Bailey, C. M.; Domaoal, R. A.; Anderson, K. S.; Jorgensen, W. L., Optimization of azoles as anti-human immunodeficiency virus agents guided by free-energy calculations. *JACS* **2008**, 130, 9492-9499.
304. Ren, P. Y., Jiao, D., Zhang, J.J., Duke, R.E., Li, G.H., Schnieders, M.J., Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential *J Comput Chem* **2009**, 30, 1701-1711.

305. Ge, X. X., Roux, B., Absolute binding free energy calculations of sparsomycin analogs to the bacterial ribosome. *J Phys Chem. B* **2010**, 114, 9525-9539.
306. Michel, J.; Essex, J. W., Hit identification and binding mode predictions by rigorous free energy simulations. *J Med Chem* **2008**, 51, 6654-6664.
307. McCammon, J. A. Computer-aided drug discovery: physics-based simulations from the molecular to the cellular level. In *Physical Biology: From Atoms to Medicine*, Zewail, A. H., Ed.; Imperial College Press: London, England, 2008, pp 401-410.

Chapter 2

Molecular Dynamics of the R175H in the Full-Length p53 Tetramer Reveal Insight into the DNA Search and Recognition Mechanism

Tavina L. Offutt, Pek U. Jeong, Özlem Demir, Rommie E. Amaro*

Department of Chemistry and Biochemistry, University of California, San Diego, 9500
Gilman Drive, La Jolla, CA 92092-0340, United States

This work is currently in preparation for submission to Nature Communications Journal.

ABSTRACT

The “guardian of the genome”, p53, functions as a tumor suppressor that responds to cell stressors such as DNA damage, hypoxia, and tumor formation by inducing cell-cycle arrest, senescence, or apoptosis. Mutation of p53 disrupts its tumor suppressor function, leading to various types of human cancers. One particular mutant, R175H, is a structural mutant that inactivates the DNA damage response pathway and acquires oncogenic functions that promotes both cancer and drug resistance. Our current work aims to understand how p53 wild type function is disrupted due to the R175H mutation. We use a series of atomistic integrative models built previously from crystal structures of the full-length p53 tetramer bound to DNA and model the R175H mutant using in silico site-directed mutagenesis. Explicitly solvated all-atom molecular dynamics (MD) simulations are performed on wild type and the R175H mutant p53. Analysis of the MD trajectories reveal insight into how wildtype p53 searches and recognizes DNA, and how this mechanism is disrupted as a result of the R175H mutation. Specifically, the optimal quaternary DNA binding mode of the DNA binding domain and how this binding mode is altered as a result of the R175H mutation in combination with zinc loss is revealed. We explain these differences in the binding modes due to differences in the dynamic characteristics of the DNA binding domain and the C-terminal domain.

INTRODUCTION

p53, commonly referred to as the “guardian of the genome,” functions as a tumor suppressor. p53 responds to various environmental stressors such as DNA damage, hypoxia, and tumor formation.¹ Once activated, p53 induces cell-cycle arrest, senescence, or apoptosis via either transcriptional or non-transcriptional pathways.^{2, 3, 4} Mutation of p53 disrupts its tumor suppressor function, leading to various types of human cancers, thereby making p53 a major drug target. Our previous computational and experimental studies have revealed stictic acid as a reactivation compound for the R175H p53 mutant.⁵ In an effort to expand on this work, we glean insight into how the R175H alters the dynamics of p53 and abrogates its DNA binding abilities. Understanding how full-length p53 (fl-p53) binds DNA and how this binding is disrupted via oncogenic mutations at an atomic level can aid in the discovery of novel reactivation compounds.

Fl-p53 contains intrinsically disordered regions and binds DNA as a homotetramer in cells.⁶ Fl-p53 comprises 393 residues that form a N-terminal domain (NTD), proline-rich domain, core DNA-binding domain (DBD), flexible linker region, tetramerization domain (TET), and a C-terminal domain (CTD).⁷ The flexible NTD is responsible for activating transcription factors. The proline-rich domain has been implicated in apoptotic activity. The DBD is involved in DNA binding, and the TET domain is crucial for tetramer formation. There remains controversy in the role of the CTD in fl-p53 DNA binding.⁸ Some studies suggest that the CTD serves as a negative regulator by blocking DBD binding to short strands of specific response elements (REs).⁹ On the other hand, other studies suggest that the CTD acts as a positive regulator of DNA binding by assisting the DBD in target site recognition in long or circular DNA.¹⁰ Our

previous computational studies of the fl-p53 bound to three different DNA response elements revealed that the CTDs approach and directly contact the DNA independent of the DNA sequence.¹¹

Due to the highly dynamic nature of p53, obtaining an experimental three-dimensional structure of fl-p53 is a challenge. The p53 DBD is the most studied due to its defined secondary and tertiary structural elements, allowing for structural characterization. Also, this region contains the majority of oncogenic mutations.¹² While these experimental structures of the DBD has provided useful information about p53 function, it is crucial to model fl-p53 in order to fully elucidate the DNA-binding mechanism under normal biological conditions. For example, how p53 searches and recognizes specific REs remains unclear. Therefore, the disruption of this search and recognition process due to p53-inactivating mutations is also not well understood.

In an effort to gain insight into the DNA binding mechanism, researchers elucidated a crystal structure of the tetrameric p53 DBD and TET with truncated linker regions bound to a short strand of DNA.¹³⁻¹⁵ In previous work, we utilized this crystal structure to build atomistic integrative models of fl-p53 bound to 3 DNA sequences (two REs and a non-specific DNA), to explore their dynamics via molecular dynamics (MD) simulations.¹¹ MD-generated ensembles agreed well with previously determined electron microscopy maps and revealed different quaternary binding modes of the fl-p53 bound to different DNA response elements. Our current study expands this work by investigating how oncogenic mutations, specifically the R175H mutation, perturbs the DNA binding interactions of fl-p53.

The R175H mutation results in inactivation of the Mre11/ATM-dependent pathway involved in DNA damage response.¹⁶ This structural mutant also results in abrogation of DNA-binding and perturbs the structure of the p53 DBD.¹⁷ It is also important to note that R175H is a gain-of-function mutant; it not only disrupts normal p53 tumor suppressor function, but also acquires alternative functions necessary for promoting cancer activities. Since this mutation is located in the L2 loop adjacent to the zinc-coordination site, zinc loss is common.^{18, 19} Zinc is important for obtaining DNA-binding specificity, and prevents aggregation of the core domain via L2 loop stabilization.²⁰ Also, in the absence of zinc, the p53 DBD is destabilized by 3.2 kcal/mol.²¹ Therefore, we hypothesize that zinc loss exacerbates the effects of the R175H mutation on the DBD. In an effort to monitor this, we use MD simulations to model the p53 DBD R175H mutant with and without zinc.

In the current study, we use a series of atomistic integrative models built previously with available crystal structures of the fl-p53 tetramer bound to DNA (Figure 2.1).¹¹ We model the R175H mutant using *in silico* site-directed mutagenesis. Explicitly solvated all-atom MD simulations are performed in duplicate on the following three fl-p53 systems: (i) wildtype, (ii) R175H mutant with zinc, and (iii) R175H mutant without zinc. Their analysis reveal differences in the conformations of the DNA binding domain and provide insight into how the R175H mutation destabilizes p53 and abrogates DNA binding.

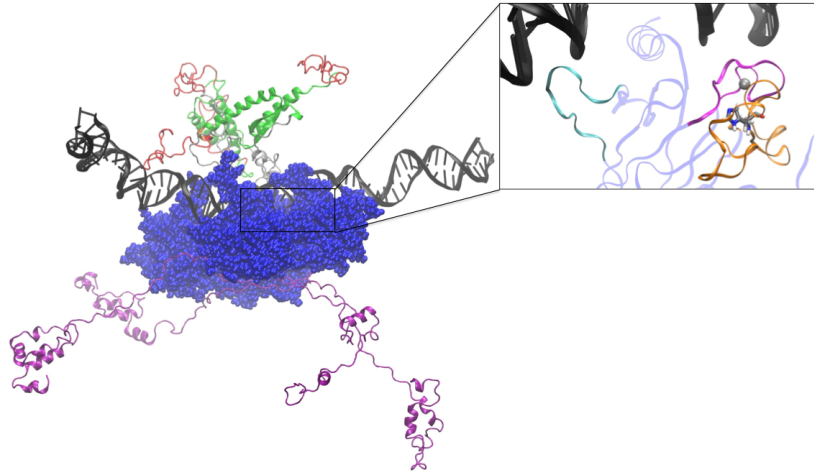


Figure 2.1: Full-length p53 System. The fl-p53 bound to DNA is shown, where the NTD, DBD, linker, TET, and CTD are colored purple, blue, silver, green, and red, respectively. DNA is depicted as a black ribbon. In the upper right panel, a monomer DBD (transparent blue ribbon) is shown with the L1, L2, and L3 loops highlighted in cyan, orange, and magenta respectively. The zinc ion is shown as a silver sphere, and the H175 residue is shown in licorice colored by atom type (C: silver, O: red, N: blue, H: white).

RESULTS

We performed 200 ns of MD simulations of all three wildtype and R175H mutant fl-p53 systems, resulting in a total simulation time of 600 ns (Table 2.S1). We explored how the R175H structural mutation with and without zinc changes the dynamic behavior of fl-p53 both globally and locally. Ultimately, we glean insight into how these changes in flexibility destabilizes the DBD and alters DNA binding in R175H p53.

Unique Binding Modes of Wildtype and R175H p53 DBD Tetramer to DNA

When projected onto 2D principal component (PC) space, the DBD tetramer for each fl-p53 system differs in DNA binding modes (Figure 2.2 and 2.S1). Interestingly, first principal motion (PC1) described a global conformational change going from an

asymmetric mode (low PC1) where monomers A and D are pushed away from the DNA and monomers B and C are close together to a symmetric binding mode (high PC1) in which all four monomers are in close proximity to DNA (Figure 2.2a). In our simulations, the wildtype and R175H with zinc tetramer systems solely sampled high PC1 values and the R175H without zinc mostly sampled low PC1 values. This means zinc loss is far more important in DNA binding failure of p53 than the R175H mutation itself.

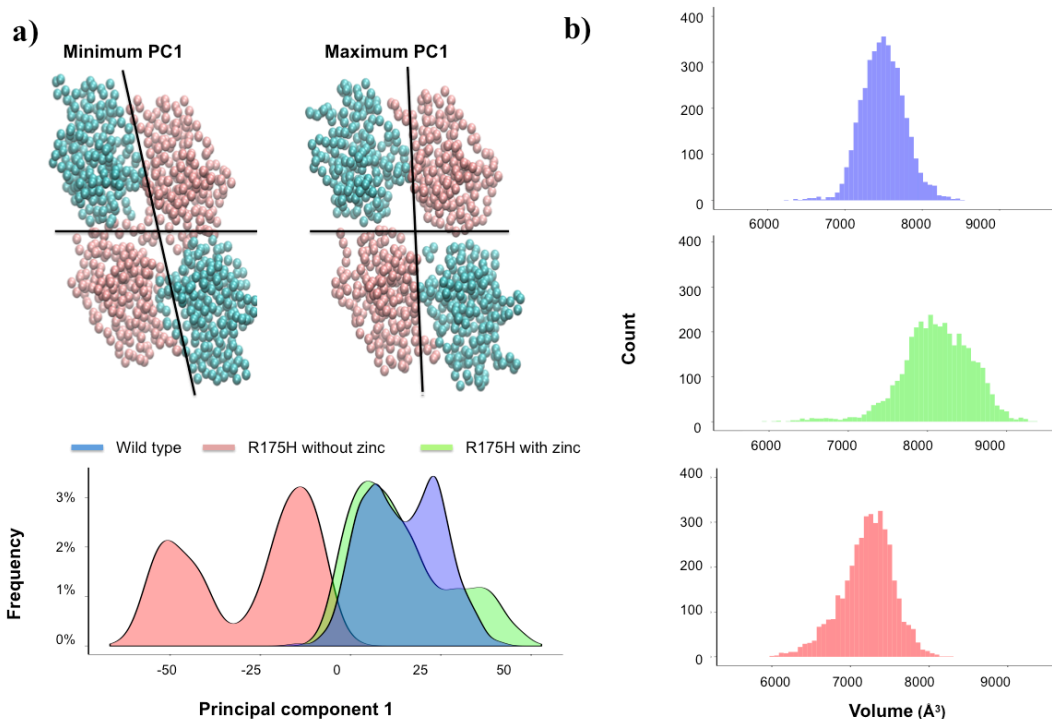


Figure 2.1: Quaternary binding mode of the DNA binding domain in wildtype and R175H p53. The eigenvalues along PC1 are shown as a density plot for each fl-p53 system (a), where the asymmetric binding mode corresponds to low PC1 (R175H without zinc) and the symmetric binding mode corresponds to maximum PC1 (wildtype and R175H with zinc). The alpha carbons of the DBD for low and high PC1 are shown as spheres, where monomers A and D are colored cyan, and monomers B and C are colored pink. A histogram plot of the DNA grip volume is shown for all three fl-p53 systems (b).

In addition to the change in symmetry of the DBD monomers, the PC1 motion also revealed differences in the DNA grip volume, which is the space in between the four p53 DBD monomers that accommodates the DNA. This volume is largest for R175H with zinc, smaller for wildtype, and smallest for R175H without zinc (Figure 2.2b). The large difference in the DNA grip volume is due to changes in the L2 loop conformations of monomers B and C. When the grip volume is small in R175H without zinc system, the L2 loops in these two monomers are stabilized via hydrogen bonding interactions (Table 2.1). In the wildtype system, one hydrogen bond is seen in the wild type p53 system in only 0.05% of the trajectory (Table 2.1), and no hydrogen bonds are observed for the R175H with zinc system. There are also differences seen in a salt-bridge near the mutation site in the L2 loop between residues Q180 and R174 (Figure 2.S2). In the wildtype system, this salt-bridge is persistent throughout the simulation, where it occurs $66.4 \pm 5.80\%$ of the time across both MD copies. However, in the mutant systems, this salt-bridge is less persistent with one exception. In the R175H with zinc system, this salt bridge forms $43.8 \pm 1.20\%$ of the time across both MD copies. For R175H without zinc, this salt-bridge only forms $47.9 \pm 43.8\%$ of the time across both MD copies; it should be noted that the large standard deviation for the R175H without zinc system is due to the fact that the disrupted salt-bridge is seen in only one MD copy.

Table 2.1: Hydrogen Bonding Interactions between L2 loops in Monomers B and C

Wildtype MD copy 1		
Monomer B L2 Residue	Monomer C L2 Residue	% interaction
194	183	0.05%
186	183	0.05%
R175H without zinc MD copy 1		
184	183	39.6%
182	184	20.3%
182	183	11.4%
185	181	8.15%
183	183	3.55%

PC2 appears to be similar to PC1, but shows more local motion of the global symmetry motion seen in PC1. In PC2, monomers A and D rotate in opposite directions (Figure 2.S1). As seen in PC1 motion, monomers B and C move closer to each other going from low PC2 to high PC2.

Each fl-p53 system sampled conformations not sampled by the other two systems, which can be defined as unique conformations. In order to further visualize the differences in these unique conformations, root-mean-square deviation (RMSD) clustering was performed on the L2 and L3 loops since these regions are local to the R175H mutation site, and the L2 loop stabilizes the L3 loop. The representative frame from the most-populated cluster was used for comparison. When clustering on the L2 loop, the differences in the L2 loop were similar to those seen in the PC1 motion, where

the L2 loops of monomers B and C formed direct contacts in the R175H without zinc system, unlike the R175H with zinc and wildtype systems (Figure 2.3a).

When clustering on the L3 loop, the conformation of R248 residue, a crucial DNA contact, in monomer C differ (Figure 2.3b). In the wildtype cluster representative, R248 sticks directly into the DNA minor groove as seen in the crystal structure (PDB 1TSR). In the R175H with zinc cluster representative frame, R248 adopts an alternate conformation that does not intercalate into the DNA minor groove. In the R175H without zinc cluster representative frame, R248 residue remains on the surface, but is completely flipped out of the minor groove. The conformation of R248 in monomers A, B, and D in the top cluster representative is similar across all three fl-p53 systems. In order to explain the differences in the quaternary binding modes, we explore the dynamics of the DBD, which will be discussed next.

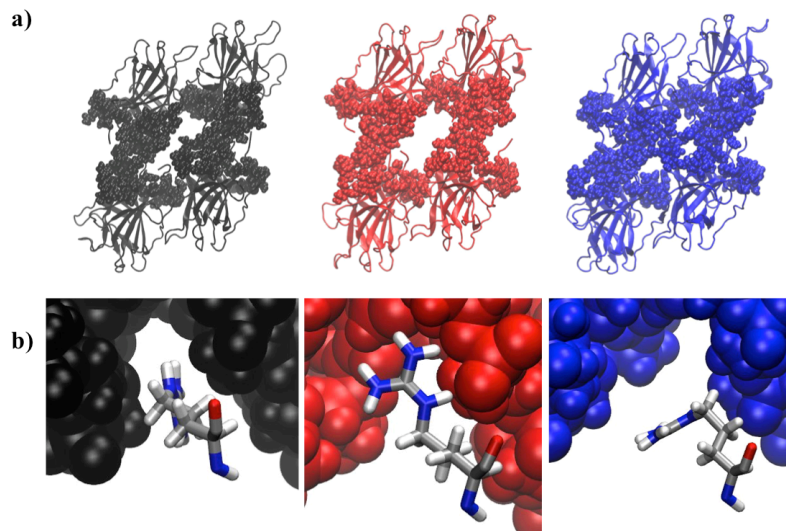


Figure 2.2: Effects of p53 R175H mutation on the L2 and L3 loops in the Unique PC Conformations. The most-populated cluster representative frame for each fl-p53 system reveals differences in the L2 loop conformation (a) and R248 residue in the L3 loop (b). In (a), the DNA binding domain for wildtype, R175H with zinc, and R175H without zinc systems are shown as black, red, and blue ribbons, respectively; the L2 loop atoms are shown as van der Waals sphere. In (b), the conformation of Arg248 is shown in sticks and colored by atom type (H: white, C: silver, N: blue, O: red). The DNA for wildtype, R175H with zinc, and R175H without zinc systems are depicted as a van der Waals sphere colored black, red, and blue, respectively.

Comparison of atomic fluctuations within the DBD

Increased fluctuations of motifs within the DBD are seen for the R175H with zinc system in monomers B, C, and D (Figure 2.S3). For the two inner monomers, B and C, the L2 and L3 loops are more dynamic in the R175H with zinc system compared to the wildtype system. Among the outer monomers, monomer D becomes more flexible at the loop between beta strands 7 and 8 in the R175H with zinc system. Surprisingly, the dynamic behavior of monomer A is similar in the R175H with zinc and wildtype systems.

Similar to the R175H with zinc system, an increase in atomic fluctuations is seen in the R175H without zinc system compared to the wildtype system (Figure 2.S4). This difference in flexibility is seen in the L2 and L3 loops for the inner monomers B and C. Interestingly, the L1 loop and H2 helix in monomer C also show greater fluctuations in the R175H without zinc system. Among the outer monomers, monomer D also shows increased flexibility at the L2 and L3 loops. In monomer A, the only RMSF variation is observed at the loop between beta strands 7 and 8.

While we don't see more flexibility in the R175H mutant systems across the entire tetramer DBD or even an entire monomer, we do observe increased fluctuations local to the R175H mutation site in monomers B, C, and D. It is striking that we do observe fluctuations distal to the mutation site as seen in monomer D in R175H with zinc system and monomer C in R175H without zinc system. In an effort to explore the global effects of this local increased flexibility, we measured and compared the solvent exposure of the DBD.

Comparison of the solvent accessible surface area of the DBD Tetramer

Protein fluorescence experimental studies have revealed differences in fluorescence spectra between wildtype and mutant p53, which is attributed to increased solvent accessibility in mutants.²¹ When comparing the SASA for the entire DBD tetramer, the SASA is higher for both R175H mutant systems than wildtype p53 as expected (Figure 2.4a). When calculating the average SASA across the entire simulation, increased solvent accessibility is seen for both R175H mutant systems ($\langle \text{SASA} \rangle_{\text{wildtype}} = 33362 \pm 761 \text{ \AA}^2$, $\langle \text{SASA} \rangle_{\text{R175H_with zinc}} = 34852 \pm 604 \text{ \AA}^2$, $\langle \text{SASA} \rangle_{\text{R175H_without zinc}} =$

33913±651 Å²), where this increase is significantly higher for the R175H with zinc system. Next, we focused on determining the solvent exposure of only the mutant epitope residues (residues 213-217) that are known to bind an antibody that selectively recognizes the mutant conformation of p53, Pab240.²² These residues are buried in the crystal structure, and must become more solvent exposed for antibody binding.

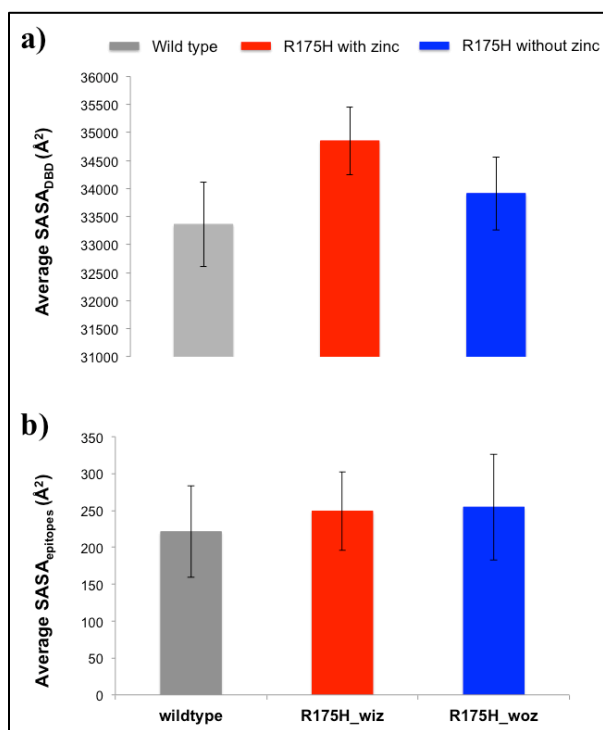


Figure 2.4: Solvent accessibility of p53 DBD. The average solvent accessible surface area of the DBD (a) and mutant epitopes that bind the mutant p53 antibody (b) are shown and reveal the R175H mutant systems have increased solvent accessibility.

Similar to the entire DBD tetramer, when we focus only on the mutant epitopes, the average SASA is higher for both R175H mutant systems compared to the wildtype (Figure 2.4b). When calculating the average SASA across all MD trajectories, increased solvent accessibility is seen for both mutant systems ($\langle \text{SASA} \rangle_{\text{wildtype}} = 221 \pm 62 \text{ \AA}^2$,

$\langle \text{SASA} \rangle_{\text{R175H_with zinc}} = 249 \pm 53 \text{ \AA}^2$, $\langle \text{SASA} \rangle_{\text{R175H_without zinc}} = 254 \pm 72 \text{ \AA}^2$), although this increase is not statistically significant. Taken together, the SASA results for the entire DBD tetramer and mutant epitopes corroborate experimental evidence that suggest that the R175H mutant form has increased solvent accessibility.

Comparison of C-terminal domain contacts with DNA

Previous computational models¹¹ have shown that the CTDs form direct DNA contacts regardless of the DNA response element. In addition, Friedler *et al.* suggest these CTD-DNA contacts are likely due to low affinity electrostatic interactions between positively charged residues in the CTD and the negatively charged DNA phosphate backbone.²³ As a result of these studies, we explore if these CTD-DNA contacts change as a result of the R175H mutation. Our MD simulations revealed there are in fact differences in the CTD-DNA contacts, with the largest change seen in the R175H without zinc system (Table 2.S2). In all three fl-p53 systems, all the positively charged CTD residues' (Lys, His, Arg) interactions with the negatively charged DNA phosphates were monitored. In monomer C, the CTD forms contacts with the DNA throughout the entire simulation across all three fl-p53 systems. These contacts are transient, in which the positive CTD residues change the phosphate groups they contact. It should be noted that the initial starting structure for the MD simulations across all three fl-p53 systems already had this CTD-DNA contact while the CTD in the other monomers start far away from the DNA. In the wildtype and R175H with zinc systems, the CTD in monomers A and D move closer to and forms DNA contacts. Remarkably, other than the CTD in monomer C, no additional CTD-DNA contacts are observed in the R175H without zinc simulations,

suggesting that the loss of zinc due to the R175H mutation disrupts CTD-DNA interactions.

Next, we closely monitored if certain positively charged CTD residues were responsible for the transient salt-bridge interactions with DNA, and identified four residues (Figures 2.5a and 2.S5). A cutoff value of 600 was chosen since this is half of the maximum number of CTD-DNA salt-bridge contacts. For both wildtype and R175H with zinc systems, CTD residues K370, R379, K381, and K382 form at least 600 DNA contacts. For R175H without zinc, only residues H368 and R379 engage in at least 600 DNA contacts. Interestingly, residues K370, K381, and K382 have reduced DNA contact when compared to wildtype and R175H with zinc. When looking at the average across all three p53 systems, K370, R379, K381, and K382 form at least half of the maximum contacts (Figures 2.5a and 2.S5).

Similarly to the CTD, we also identified the DNA residues that the CTD residues interact with, and found that salt bridges formed both inside and outside the response element (Figures 2.5b and 2.S6). Across all three p53 systems, only a few contacts are made within the response element (residues 1599-1602, 1665, 1681-1684) (black box in Figures 2.5b and S6). Majority of the CTD-DNA interactions occur outside the DNA response element, interacting with DNA nucleotides up 10 base pairs to the right and 14 base pairs to the left away from the response element (Figure 2.5b). The CTD searches furthest along the DNA in the R175H mutant systems (Figure 2.S6).

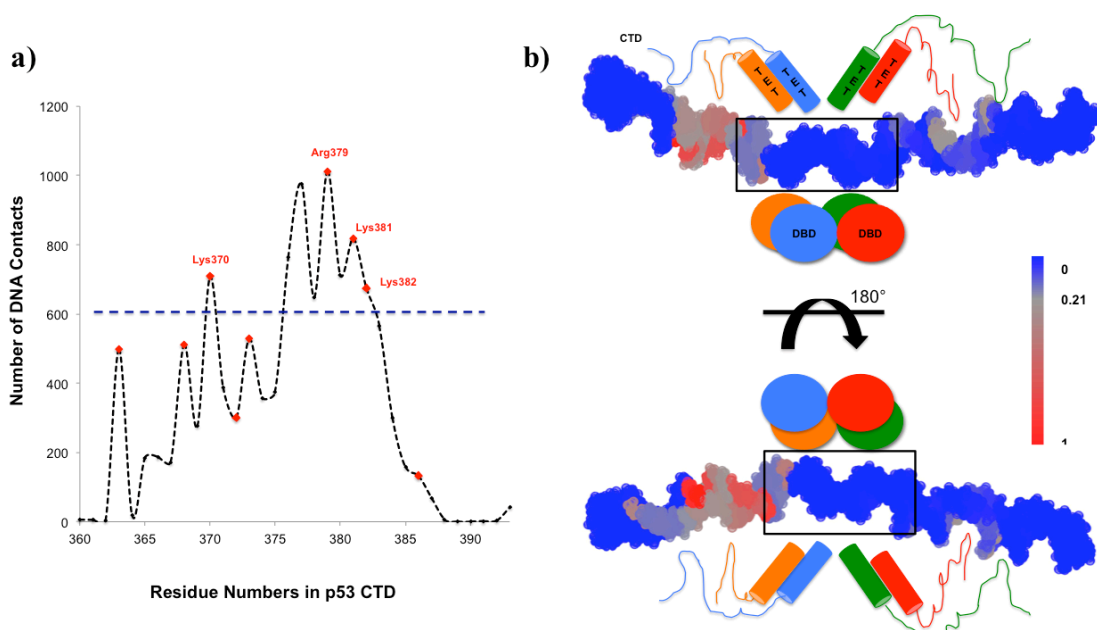


Figure 2.5: Footprint analysis of the CTD-DNA contact residues averaged across all three p53 systems. The residues in the CTD that come within 3.5Å of the DNA were averaged across each monomer and all three p53 systems (a). The positively charged CTD residues are highlighted in red. A value cutoff for the number of DNA contacts formed of 600 is selected (dashed blue line) since it is at least half of the maximum CTD-DNA contacts. The four positively charged CTD residues that meet this cutoff are labeled. The same footprint analysis is done from the perspective of the DNA, in which the number of CTD contacts is mapped onto the DNA (b). The number of CTD contacts are normalized, ranging from 0 (no CTD contacts) to 1 (maximum number of CTD contacts). A cartoon of the DBD, TET, and CTD domains are depicted to highlight the orientation of the DNA, and the DNA response element is highlighted with a black box.

L1/S3 Pocket Opening

In an earlier study, MD simulations of wildtype and various p53 mutant DBDs revealed a druggable L1/S3 pocket.⁵ Using several geometric criteria as a filter for determining the pocket-open state, the pocket was found to be open only about 6% of the time in a 30 ns simulation. In the same study, virtual screening against the L1/S3 pocket in MD-generated conformations revealed a novel R175H reactivation compound, stictic acid.

In the current study, we compute the pocket opening of R175H mutants under normal physiological conditions using the same geometric criteria outlined in Ref 5 (Table 2.S3). For the inner monomers B and C, the L1/S3 pocket is open only 4% to 13% of the time for wildtype, 7% to 24% for R175H with zinc, and 4% to 15% for R175H without zinc. The pocket is open for majority of the simulation in the outer monomers, A and D, for all three systems with two exceptions (24% to 95% for wildtype, 67% to 96% for R175H with zinc, and 21% to 96% for R175H without zinc). These results are promising because they show that in the R175H fl-p53 mutant, the L1/S3 pocket is in fact open and available for reactivation molecules to bind in restoring p53 wildtype function.

DISCUSSION

We report here three different integrative atomistic models of the wildtype and R175H mutant fl-p53 tetramer bound to the p21 response element and their dynamics via MD simulations. Experimental studies have shown that the R175H mutation accelerates the rate of zinc loss in the DBD.^{18, 19} However, the timescale of zinc loss remains unclear. Therefore, we model the R175H mutation in both the presence and absence of zinc. There are no available experimental structures of the R175H mutant even in the DBD of p53 due to the denaturing effects of the mutation. To our knowledge, the dynamics of the R175H mutation in the fl-p53 and its effects on DNA binding have not been explored.

R175H shifts the DBD Quaternary Binding Mode to Resemble p53 binding to Non-specific DNA

In order to provide context for the DBD quaternary binding modes revealed in the current work, we discuss our previous computational study where fl-p53 was bound to three different DNA response elements (p21 RE, puma RE, and non-specific DNA).¹¹ The binding affinity for fl-p53 for known REs, p21 and puma for example, are an order of magnitude lower than the K_D values of non-specific DNAs under physiological conditions.²⁴ Previous work revealed different quaternary binding modes for fl-p53 bound to different types of DNA in an effort to explain the different K_D values.

The PC1 motion revealed in the current work is the same global motion that was seen as PC2 in our previous computational studies.¹¹ In the previous work, the p21-bound wildtype p53 tetramer system sampled only high PC2 values corresponding to a symmetric binding mode. The same binding mode is seen in our current wild type simulations (Figure 2.2). The only difference between the p21-bound p53 tetramer in the previous work and our current p21-bound p53 tetramer system (referred to as wildtype) is the starting conformation for the MD simulations. In the previous model, the fl-p53 started from a conformation where the CTDs and NTDs were extended, and then relaxed to a more compact structure where the CTDs moved closer to the DNA.¹¹ Our current wildtype system started from the more compact fl-p53 structure. Our current computational models with additional MD sampling corroborate the results from previous studies that a symmetric DBD tetramer is ideal for p53 binding to known DNA response elements.

The PC motions of the quaternary binding modes of the mutant fl-p53 systems suggest how the R175H disrupts DNA binding. In the previous computational models, the p53 tetramer bound to nonspecific-DNA solely sampled the asymmetric binding mode.¹¹ Remarkably, the R175H mutant without zinc system mostly samples this same asymmetric quaternary binding mode, suggesting that the R175H mutant disrupts DNA binding by shifting the cooperative binding mode of the DBD tetramer to resemble non-specific DNA bound conformation.

As for the DNA grip volume change revealed in the PC motions, there are both similarities and differences when compared to the previous computational studies (Figure 2.2). The small grip volume seen in the R175H without zinc system is comparable to that seen for the non-specific DNA bound DBD tetramer. Also, as seen in the previous work, the wildtype system does have a larger DNA grip volume (ranging between 8000 and 9000 Å³) to accommodate the DNA than the DBD tetramer bound to non-specific DNA and the R175H without zinc mutant. Surprisingly, the R175H with zinc system has the largest DNA grip volume, which may be too loose to bind the DNA tightly.

Local and Global Increased Flexibility in R175H Mutant DBD Disrupt Crucial DBD-DNA Contacts

It is interesting that the motifs within the DBD with increased flexibility in the R175H mutant systems are regions known to make important DNA contacts either directly or indirectly (Figures 2.S3 and 2.S4). In the L2 loop of wildtype p53, R175 forms a salt-bridge with D184, which is thought to aid in the L2 loop stabilization as seen in various crystal structures of the wildtype DBD.^{25, 26} Therefore, the R175H mutation

disrupts this R175-D184 salt-bridge, and we also identify another salt bridge, Q180-R174, that is persistent in the wildtype simulations and disrupted in both R175H mutant simulations. Disruptions of these two salt-bridges increase the flexibility of the L2 loop (Figures 2.S3 and 2.S4). This increased flexibility in the L2 loop destabilizes the L3 loop, which does make direct contacts with the DNA via two residues, S241 and R248. Although we do not see any changes in S241-DNA contacts, we do see a difference in the R248-DNA contact, where the R248 doesn't intercalate the DNA minor groove in the R175H mutant system but does in the wildtype system (Figure 2.3b). R248 is thought to play a critical role in DNA binding because it is the most frequently mutated p53 residue in human cancers.²⁷⁻²⁹ Therefore, the loss of this DNA contact may impede DNA binding of the R175H mutant. Even though this difference in R248 conformation is only observed in one DBD monomer (monomer C), experimental studies have shown that a heterotetramer with only one mutant p53 monomer is enough to shift the wildtype p53 to resemble a mutant conformation and disrupt DNA binding.³⁰⁻³² In addition to the increased flexibility of motifs local to the R175H site, it is noteworthy that we see the effect of the mutation on long-range motions such as the H2 helix, L1 loop, and S7/S8 loop, especially in such short MD sampling (Figures 2.S3 and 2.S4). With our limited computational sampling, we are already beginning to see the destabilization of the DBD both locally at the R175H mutation site and globally.

R175H Mutation and Zinc Loss Together Disrupt C-terminal interactions with DNA

The role of the CTD in DNA binding remains controversial. Previous studies have suggested three possible theories as to how the CTD of p53 regulates sequence-

specific DNA binding.⁸ One theory suggests that the DBD tetramers only bind the DNA when it undergoes a conformational change induced by chemical modification (acetylation or phosphorylation) or protein binding of the C-terminus. Another hypothesis suggests that CTD binding to the DNA prevents the DBD tetramers from binding, and chemical modification of CTD disrupts DNA binding, thereby allowing the DBD to bind DNA. Both of these theories suggest that the CTD functions as a negative regulator for p53 binding. The third theory implies that the CTD acts as a positive regulator for DBD binding; our previous and current computational models support this theory.

In our simulations within this current study, we see transient CTD interactions with DNA (Table 2.S2). These observations are in agreement with experimental studies that reveal that the CTD forms sequence-independent contacts with DNA.^{33,34} It is particularly interesting that the CTD monomers (A and D) that were extended far away from the DNA move closer to the DNA and form direct contacts in wildtype and R175H with zinc simulations centered on CTD residues K370, R379, K381, and K382. This direct CTD-DNA contact was also seen in our previous computational models where the CTDs started from an extended conformation far from the DNA; in every simulation regardless of the DNA response element, the CTD moved closer to and directly contacted the DNA, suggesting that the CTD assists the DBD in binding DNA.¹¹ In the R175H without zinc system, the only CTD monomer that forms DNA contacts is monomer C since the starting structure involved these contacts. Unlike the other 2 systems, no other CTDs approach DNA to form contacts, and reduced DNA contacts are seen for residues

K370, K381, and K382. Taken together, the combination of R175H mutation and zinc loss disrupts CTD-DNA contacts.

CONCLUSION

In this study, we used MD simulations to explore the effects of the R175H cancer mutation on the dynamic characteristics of p53, and how these changes disrupt DNA binding. Results reveal increased flexibility of motifs within the DBD both local and distal to the R175H mutation site. Interestingly, these motifs are regions that form important DNA contacts. The increased dynamics disrupt the DBD from adopting an optimal DNA binding mode and alters the CTD-DNA contacts. Taken together our mutant models in the current work and previous models with fl-p53 bound to different DNA response elements allow us to glean insight into the DNA search and recognition mechanism even with limited MD sampling. Under normal wildtype p53 conditions, the DBD adopts a symmetric binding mode around the DNA, and the CTD centered on three Lys residues and 1 Arg residue aids in DNA binding by forming contacts with the DNA mostly outside the response element region. When p53 has the R175H cancer mutation in combination with zinc loss, the DBDs shift to an asymmetric binding mode around the DNA, and the CTD-DNA contacts are disrupted. The results of our computational models support our hypothesis that zinc loss exacerbates the effects of the R175H mutation in destabilizing the DBD and abrogating DNA binding.

MATERIALS AND METHODS

Construction of Models

The full-length p53 system was generated as described in previous work.¹¹ The last frame of the previously simulated wild type (WT) p53 system bound to the p21 response element was selected for the current study, in which the fl-p53 fully relaxed into a compact structure. All ions and water molecules were removed with the exception of waters within 4 Å of the protein. Three systems were built from the previously simulated p53-p21 system: (i) wildtype, (ii) R175H with zinc, and (iii) R175H without zinc. The R175H mutation was modeled using *in silico* site-directed mutagenesis in all four monomers. Zinc's tetrahedral geometry was modeled using the dummy cationic atom in the wild type and R175H mutant with zinc systems.³⁵ The coordinating cysteine and histidine residues were deprotonated, bearing a negative charge to model zinc-protein coordination. In the R175H mutant without zinc, the correct protonation states for cysteine and histidine residues were determined in the absence of the zinc ion using the online PDB2PQR webserver.³⁶

After all three models were built, sodium ions were added to neutralize each system. Using the TIP3P water model,³⁷ the systems were solvated in a 226 X 193 X 234 box (10 Å in the x-, y-, and z- direction). Each system consisted of ~960,000 atoms and was built using the Amber FF14SB force field.³⁸

Molecular Dynamics Simulations

All-atom explicit-solvent molecular dynamics (MD) simulations were performed for the four systems using NAMD2.12.³⁹ The general MD workflow consisted of three stages: minimization, equilibration, and production. The prepared systems were minimized in five steps as follows: (i) minimization of the protons while restraining the

protein, DNA, and solvent for 2000 steps; (ii) minimization of protons, water and ions, while restraining the DNA and protein for 2000 steps; (iii) minimization of the protein and DNA side chains only while restraining the backbone and zinc ion for 2000 steps; (iv) minimization of the zinc coordinating residues only while restraining the non-zinc coordinating protein residues and DNA for 10000 steps; (v) minimization of all atoms in system for 20000 steps. The non-bonded energy was calculated every minimization step. Long-range interactions were calculated using the Particle Mesh Ewald method with a cut-off distance of 10Å.⁴⁰ At 8Å, a switching function was applied to improve energy conservation.

The minimized systems were then equilibrated using the NVT ensemble in four steps. All heavy atoms were restrained starting from a weight of 4kcal/mol and reduced gradually to 1kcal/mol. The systems were heated to a temperature of 310K and maintained with Langevin dynamics with a damping coefficient of 5 picoseconds/terahertz. Following equilibration, an NPT ensemble was performed with no positional constraints. A Langevin piston barostat was used to hold the pressure constant at 1 atm with an oscillation period of 100 femtoseconds (fs) and a damping time scale of 50 fs. Two production runs were performed, resulting in a total simulation time of 200ns for each system. Every 5th frame for the simulation was saved and used for analysis.

Principal Components Analysis

The Amber tools cpptraj package was used to perform principal component analysis on the DNA binding domain of p53.^{41, 42} The trajectories across all three systems were concatenated and aligned on the α -carbons in the DBD (residues 89 to 291)

to remove translation and rotation. The variances of the α -carbon coordinates were determined using the starting conformation for MD as a reference. These variances were used to generate a covariance matrix, A , which was then diagonalized to reveal eigenvectors 1 and 2. Calculation of the covariance matrix A was conducted as follows, yielding the eigenvalues, λ : $A\mu = \lambda\mu$. These eigenvalues along eigenvectors 1 and 2 were plotted using gnuplot,⁴³ and used to compare the conformational space of the DBD in the wild type and mutant p53 MD simulations. Pseudotrajectories were generated to visualize the motion of each eigenvector.

RMSD Clustering Analysis

MD frames unique to each p53 system were extracted and clustered in order to further visualize and analyze PCA results. ‘Unique frames’ are frames that when projected into PC space, do not overlap with conformations from the other p53 systems. The unique frames were extracted using a python script that selected frames within certain eigenvalue cutoffs (Figure 2.S1), resulting in 650 frames for clustering per system. The extracted frames were aligned to the starting structure for MD on the α -carbons in the DBD. Using the Gromos algorithm,^{44, 45} pairwise root-mean-square deviations were calculated and used for clustering the heavy atoms in the L2 and L3 loops L3 loop. The following cutoffs were selected for the L2 and L3 loops respectively: 1.1 Å and 0.9 Å. The top cluster representative for each p53 system was used to visualize differences in the three DBD motifs.

Root-mean-square fluctuation Analysis

The root-mean-square fluctuations (RMSF) of the DBD residues for each monomer were calculated using cpptraj.^{41,42} RMSF is a measure of how a system fluctuates about a well-defined average position. For each p53 system, the MD trajectories were aligned to the starting MD structure using the DBD backbone atoms (N,C α ,C,O) for each monomer, and an average structure was calculated. Using this average structure as the reference, the RMSF of the DBD residues (using only the atoms that make up the backbone) was calculated and plotted using gnuplot.⁴³

Volume Calculation

The grab volume between the four DBD monomers that accommodates DNA binding was calculated using POVME 2.0.⁴⁶ The visual molecular dynamics (VMD) program⁴⁷ was used to generate an inclusion sphere that centered at Cartesian coordinates (128, 135, 115) with a radius of 17 Å, which fully engulfed the volume between the four monomers. A seed was planted in the center of the sphere and extended for 4 Å. POVME 2.0 calculated the grab volume starting from the seed and continued until it reached the inclusion region boundary. The volume was calculated for every MD snapshot, and the volume distribution was plotted as a histogram using the R program.⁴⁸

Solvent Accessible Surface Area Analysis

The solvent accessible surface area (SASA), or the exposed areas of atoms, of the tetramer core domain was measured for each p53 system using cpptraj.⁴² The SASA is described as rolling a solvent sphere over the van der Waals surface of a protein. The SASA was calculated in Å² using the linear combinations of pairwise overlaps (LCPO)

algorithm.⁴⁹ In the LCPO method, each atom in the protein is represented as a hard sphere. The SASA of each atom sphere was calculated as the difference between the surface area of the atom and the area of atom overlap.

Salt-Bridge Interactions between C-terminal domain and DNA

In order to investigate the C-terminal domain (CTD) - DNA contacts, we measured salt bridges between the positive residues on fl-p53 CTD and the negative DNA phosphate atoms. First, we identified all fl-p53 CTD Lys, Arg, and His residues that came within 5 Å of the DNA at any point of the MD simulation using a tool command language (tcl) script executed in VMD.⁴⁷ The trajectories were then loaded into VMD⁴⁷ and visual inspection was used to identify salt bridges between the selected Lys/Arg/His residues and DNA phosphate atoms. The distance between the positive nitrogen atoms and negative DNA phosphate oxygen atom throughout the MD trajectory were manually extracted. A python script was used to calculate the percent of the salt-bridge interaction using a distance cutoff of 3.5 Å.⁵⁰

L1/S3 Pocket Open Ratio Calculation

In order to calculate the percentage of the L1/S3 pocket opening, the same distance and angle criteria outlined in Wassman *et al.*⁵ was used. First, we used cpptraj⁴² to define the four distances and one dihedral angle associated with the L1/S3 pocket that served as input for the calculation. Next, using an in-house python script, the frames that satisfied the distance and angle criteria were identified, and used in calculating the percentage of the time the pocket was open.

REFERENCES

1. Freed-Pastor, W. A.; Prives, C., Mutant p53: one name, many proteins. *Genes Dev* **2012**, 26, 1268-86.
2. Green, D. R.; Kroemer, G., Cytoplasmic functions of the tumour suppressor p53. *Nature* **2009**, 458, 1127-1130.
3. Li, T. Y.; Kon, N.; Jiang, L.; Tan, M. J.; Ludwig, T.; Zhao, Y. M.; Baer, R.; Gu, W., Tumor Suppression in the Absence of p53-Mediated Cell-Cycle Arrest, Apoptosis, and Senescence. *Cell* **2012**, 149, 1269-1283.
4. Valente, L. J.; Gray, D. H. D.; Michalak, E. M.; Pinon-Hofbauer, J.; Egle, A.; Scott, C. L.; Janic, A.; Strasser, A., p53 Efficiently Suppresses Tumor Development in the Complete Absence of Its Cell-Cycle Inhibitory and Proapoptotic Effectors p21, Puma, and Noxa. *Cell Rep* **2013**, 3, 1339-1345.
5. Wassman, C. D.; Baronio, R.; Demir, O.; Wallentine, B. D.; Chen, C. K.; Hall, L. V.; Salehi, F.; Lin, D. W.; Chung, B. P.; Hatfield, G. W.; Richard Chamberlin, A.; Luecke, H.; Lathrop, R. H.; Kaiser, P.; Amaro, R. E., Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nat Commun* **2013**, 4, 1407.
6. Tidow, H.; Melero, R.; Mylonas, E.; Freund, S. M.; Grossmann, J. G.; Carazo, J. M.; Svergun, D. I.; Valle, M.; Fersht, A. R., Quaternary structures of tumor suppressor p53 and a specific p53 DNA complex. *Proc Natl Acad Sci USA* **2007**, 104, 12324-9.
7. Hupp, T. R., Regulation of p53 protein function through alterations in protein-folding pathways. *Cell Mol Life Sci* **1999**, 55, 88-95.
8. Ahn, J.; Prives, C., The C-terminus of p53: the more you learn the less you know. *Nat Struct Biol* **2001**, 8, 730-732.
9. Hupp, T. R.; Meek, D. W.; Midgley, C. A.; Lane, D. P., Regulation of the Specific DNA-Binding Function of P53. *Cell* **1992**, 71, 875-886.
10. McKinney, K.; Mattia, M.; Gottifredi, V.; Prives, C., p53 linear diffusion along DNA requires its C terminus. *Mol Cell* **2004**, 16, 413-424.
11. Demir, O.; Jeong, P. U.; Amaro, R. E., Full-length p53 tetramer bound to DNA and its quaternary dynamics. *Oncogene* **2017**, 36, 1451-1460.
12. Lukman, S.; Lane, D. P.; Verma, C. S., Mapping the Structural and Dynamical Features of Multiple p53 DNA Binding Domains: Insights into Loop 1 Intrinsic Dynamics. *Plos One* **2013**, 8.

13. Emamzadah, S.; Tropia, L.; Halazonetis, T. D., Crystal structure of a multidomain human p53 tetramer bound to the natural CDKN1A (p21) p53-response element. *Mol Cancer Res* **2011**, 9, 1493-9.
14. Petty, T. J.; Emamzadah, S.; Costantino, L.; Petkova, I.; Stavridi, E. S.; Saven, J. G.; Vauthey, E.; Halazonetis, T. D., An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. *EMBO J* **2011**, 30, 2167-76.
15. Emamzadah S, T. L., Vincenti I, Falquet B, Halazonetis TD Reversal of the DNA-binding-induced loop L1 conformational switch in an engineered human p53 protein. *J Mol Biol* **2014**, 426, 936-944.
16. Xu, Y., DNA damage: a trigger of innate immunity but a requirement for adaptive immune homeostasis. *Nat Rev Immunol* **2006**, 6, 261-270.
17. Sigal, A.; Rotter, V., Oncogenic mutations of the p53 tumor suppressor: the demons of the guardian of the genome. *Cancer Res* **2000**, 60, 6788-93.
18. Butler, J. S.; Loh, S. N., Structure, function, and aggregation of the zinc-free form of the p53 DNA binding domain. *Biochemistry* **2003**, 42, 2396-403.
19. Joerger, A. C.; Fersht, A.R., Structure-function-rescue: the diverse nature of common p53 cancer mutants. *Oncogene* **2007**, 26, 2226-2242.
20. Duan, J.; Nilsson, L., Effect of Zn²⁺ on DNA recognition and stability of the p53 DNA-binding domain. *Biochemistry* **2006**, 45, 7483-92.
21. Bullock, A. N.; Henckel, J.; Fersht, A. R., Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene* **2000**, 19, 1245-56.
22. Gannon, J. V.; Greaves, R.; Iggo, R.; Lane, D. P., Activating mutations in p53 produce a common conformational effect. A monoclonal antibody specific for the mutant form. *EMBO J* **1990**, 9, 1595-602.
23. Friedler, A.; Veprintsev, D. B.; Freund, S. M.; von Glos, K. I.; Fersht, A. R., Modulation of binding of DNA to the C-terminal domain of p53 by acetylation. *Structure* **2005**, 13, 629-36.
24. Weinberg, R. L.; Veprintsev, D. B.; Fersht, A. R., Cooperative binding of tetrameric p53 to DNA. *J Mol Biol* **2004**, 341, 1145-59.

25. Kitayner, M.; Rozenberg, H.; Kessler, N.; Rabinovich, D.; Shaulov, L.; Haran, T. E.; Shaked, Z., Structural basis of DNA recognition by p53 tetramers. *Mol Cell* **2006**, 22, 741-53.
26. Wang, Y.; Rosengarth, A.; Luecke, H., Structure of the human p53 core domain in the absence of DNA. *Acta Crystallogr D Struct Biol* **2007**, 63, 276-81.
27. Takahashi, T.; Nau, M. M.; Chiba, I.; Birrer, M. J.; Rosenberg, R. K.; Vinocour, M.; Levitt, M.; Pass, H.; Gazdar, A. F.; Minna, J. D., p53: a frequent target for genetic abnormalities in lung cancer. *Science* **1989**, 246, 491-4.
28. Hollstein, M.; Sidransky, D.; Vogelstein, B.; Harris, C. C., p53 mutations in human cancers. *Science* **1991**, 253, 49-53.
29. Nigro, J. M.; Baker, S. J.; Preisinger, A. C.; Jessup, J. M.; Hostetter, R.; Cleary, K.; Bigner, S. H.; Davidson, N.; Baylin, S.; Devilee, P.; et al., Mutations in the p53 gene occur in diverse human tumour types. *Nature* **1989**, 342, 705-8.
30. Chene, P., In vitro analysis of the dominant negative effect of p53 mutants. *J Mol Biol* **1998**, 281, 205-9.
31. Milner, J.; Medcalf, E. A., Cotranslation of activated mutant p53 with wild type drives the wild-type p53 protein into the mutant conformation. *Cell* **1991**, 65, 765-74.
32. Milner, J.; Medcalf, E. A.; Cook, A. C., Tumor suppressor p53: analysis of wild-type and mutant p53 complexes. *Mol Cell Biol* **1991**, 11, 12-9.
33. Halazonetis, T. D.; Kandil, A. N., Conformational shifts propagate from the oligomerization domain of p53 to its tetrameric DNA binding domain and restore DNA binding to select p53 mutants. *EMBO J* **1993**, 12, 5057-64.
34. Oberosler, P.; Hloch, P.; Ramsperger, U.; Stahl, H., p53-catalyzed annealing of complementary single-stranded nucleic acids. *EMBO J* **1993**, 12, 2389-96.
35. Pang, Y. P., Novel zinc protein molecular dynamics simulations: Steps toward antiangiogenesis for cancer treatment. *J Mol Model* **1999**, 5, 196-202.
36. Dolinsky TJ, N. J., McCammon JA, Baker NA, PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculation. *Nucleic Acids Res* **2004**, 32, W665-W667.
37. Jorgensen WL, C. J., D. MJ, Impey RW, Klein ML, Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **1983**, 79, 926-935.

38. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015**, 11, 3696-713.
39. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *J Comput Chem* **2005**, 26, 1781-802.
40. Darden T, P. L., Li L, Pedersen L, New tricks for models from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure* **1999**, 7, R55-60.
41. D.A. Case, J. T. B., R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman, AMBER 2015. **2015**, University of California, San Francisco.
42. Roe, D. R.; Cheatham, T. E., 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013**, 9, 3084-95.
43. Janert, P. K., *Gnuplot in action : understanding data with graphs*. Manning Publications: Greenwich, Conn., 2010; p xxxi, 360 p.
44. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., GROMACS: fast, flexible, and free. *J Comput Chem* **2005**, 26, 1701-18.
45. Xavier Daura, K. G., Bernhard Jaun, Dieter Seebach, Wildfred F. van Gusteren, Alan E. Mark, Peptide Folding: When Simulataion Meets Experiment. *Angewandte Chemie* **199**, 38, 236-240.
46. Durrant, J. D.; Votapka, L.; Sorensen, J.; Amaro, R. E., POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J Chem Theory Comput* **2014**, 10, 5047-5056.
47. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* **1996**, 14, 33-8, 27-8.
48. H, W. Ggplot2 elegant graphs for data analysis. In *Use R*; Springer: New York, 2009, pp viii, 212 pages.

49. Weiser, J.; Shenkin, P. S.; Still, W. C., Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J Comput Chem* **1999**, 20, 217-230.
50. Lee, H.; Mok, K. H.; Muhandiram, R.; Park, K. H.; Suk, J. E.; Kim, D. H.; Chang, J.; Sung, Y. C.; Choi, K. Y.; Han, K. H., Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J Biol Chem* **2000**, 275, 29426-32.

Supporting Information

Table 2.S1: Summary of Simulated Model Systems

Model	Total Atoms Simulated	Total Simulation Time
Wildtype	967,380	100ns x 2copies = 200ns
R175H with zinc	966,873	100ns x 2copies = 200ns
R175H without zinc	954,048	100ns x 2copies = 200ns

Table 2.S2: Salt Bridge Footprint Analysis of CTD-DNA

p53 Residue	DNA counterpart	% interaction	p53 Residue	DNA counterpart	% interaction
Wild type MD copy1			Wild type MD Copy2		
372	DC1665 DC1660	1.65% 0.200%	363	DA1599 DA1600	8.65% 10.75%
373	DC1665	0.800%	368	DT1686	7.85%
363	DA1600 DA1602	6.90% 1.00%	370	DT1686 DT1687	63.0% 16.75%
370	DT1686 DT1687	44.7% 42.9%	373	DC1592 DT1686 DT1687	26.0% 3.25% 5.30%
373	DT1686 DT1687	12.5% 22.1%	379	DA1689 DA1690	54.8% 27.7%
379	DA1687 DA1688 DA1689 DT1690	37.0% 36.8% 17.1% 14.60%	381	DA1688 DA1689	25.8% 40.7%
381	DC1589 DA1590 DA1689	10.5% 2.25% 19.6%	382	DT1687 DT1688	5.4% 90.9%
382	DA1590 DT1687 DT1688 DT1689	3.20% 10.40% 47.3% 38.8%	368	DC1587	0.350%
368	DC1589 DC1590	15.2% 16.4%			

Table 2.S2 con't: Salt Bridge Footprint Analysis of CTD-DNA

p53 Residue	DNA counterpart	% interaction	p53 Residue	DNA counterpart	% interaction
R175H with zinc MD Copy 1			R175H with zinc MD Copy 2		
363	DG1654 DG1655 DG1656	1.35% 12.5% 1.25%	370	DC1665	1.25%
370	DA1622 DA1623	8.90% 10.7%	372	DC1665 DT1664	3.55% 3.00%
372	DA1622 DA1623 DA1624 DA1625 DG1654 DA1663	1.50% 5.00% 19.2% 4.85% 2.70% 2.35%	373	DT1664 DC1665 DA1666	4.70% 10.4% 2.15%
373	DA1623 DA1624 DA1653 DG1654	4.20% 1.00% 3.85% 4.40%	363	DA1601	14.8%
379	DG1654 DG1655	10.0% 33.5%	365	DA1600 DA1601	1.00% 2.25%
370	DT1686 DT1687	9.05% 1.20%	370	DT1686 DT1687	2.60% 89.7%
373	DT1685 DT1686 DT1687	28.9% 40.2% 33.1%	372	DC1589 DA1590	38.1% 45.8%
379	DA1689 DT1690	86.4% 44.9%	373	DT1686 DT1687	7.45% 19.4%
380	DA1591	5.35%	379	DA1689 DT1690	29.0% 30.0%
381	DA1689	59.9%	380	DA1590	51.1%
382	DT1687 DT1688	36.8% 73.0%	381	DA1688 DA1689	4.70% 62.4%
368	DC1589	0.250%			

Table 2.S2 con't: Salt Bridge Footprint Analysis of CTD-DNA

p53 Residue	DNA counterpart	% interaction	p53 Residue	DNA counterpart	% interaction
R175H without zinc MD copy 1			R175H without zinc MD copy 2		
368	DT1682	4.95%	365	DA1596 DA1597	16.0% 36.0%
370	DT1682 DT1683	88.5% 9.10%	368	DA1597	25.4%
372	DA1586 DA1685	2.55% 5.90%	370	DT1682 DT1683	17.2% 75.6%
373	DC1588 DT1682 DT1683	17.5% 7.35% 7.30%	373	DT1682 DT1683	75.9% 28.4%
379	DA1685 DT1686	15.2% 23.9%	379	DA1685 DA1686	21.3% 26.3%
380	DC1585 DC1586	8.80% 1.30%	380	DA1585	3.40%
381	DT1684 DA1685 DT1686	7.40% 59.5% 1.25%	381	DT1684 DA1685	1.30% 70.4%
382	DT1683 DT1684 DT1685	50.0% 46.9% 1.45%			

The residues highlighted in pink, orange, and blue are from monomers A, C, and D respectively. The DNA bases in the response element region are boldened.

Table 2.S3: L1/S3 Pocket Open Ratios in MD Simulations

Monomer A			
	Wildtype	R175H with zinc	R175H without zinc
Copy 1	24.8%	96.2%	21.2%
Copy 2	8.70%	77.4%	53.5%
Monomer B			
Copy 1	4.70%	7.55%	12.4%
Copy 2	9.60%	10.2%	4.65%
Monomer C			
Copy 1	12.0%	12.4%	10.4%
Copy 2	13.1%	24.6%	15.0%
Monomer D			
Copy 1	93.6%	80.4%	96.0%
Copy 2	95.4%	67.4%	69.4%

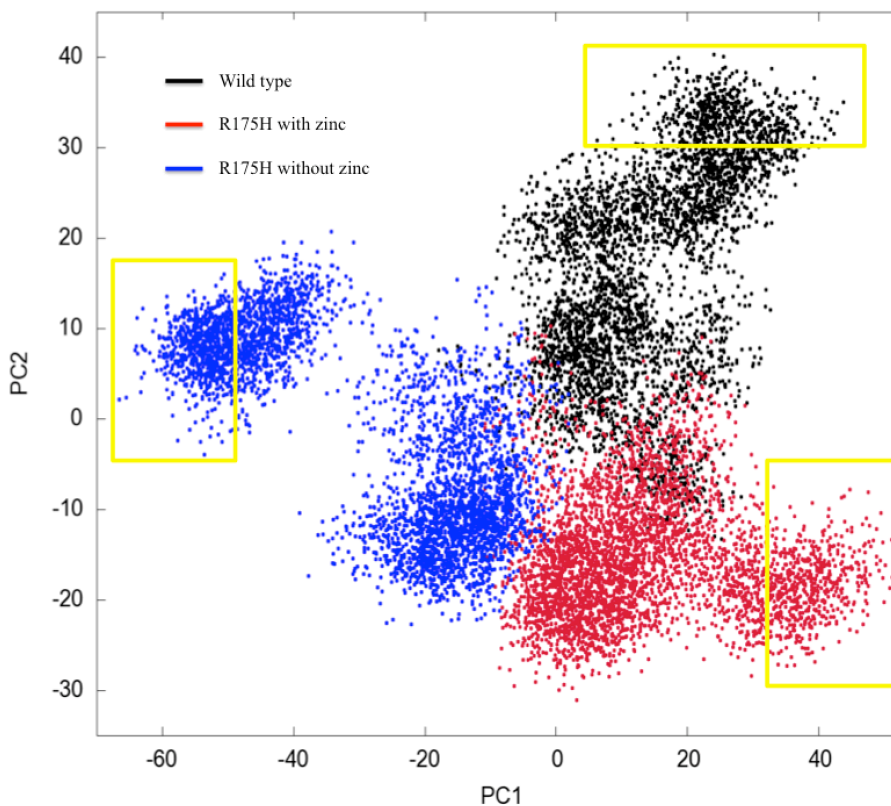


Figure 2.S1: Comparison between DNA binding modes of DBD. The eigenvalues for principal components 1 and 2 of the DBD for all three fl-p53 systems are shown. The boxed frames in yellow are the unique frames selected and extracted for RMSD clustering in order to visualize differences in the L2 and L3 loops.

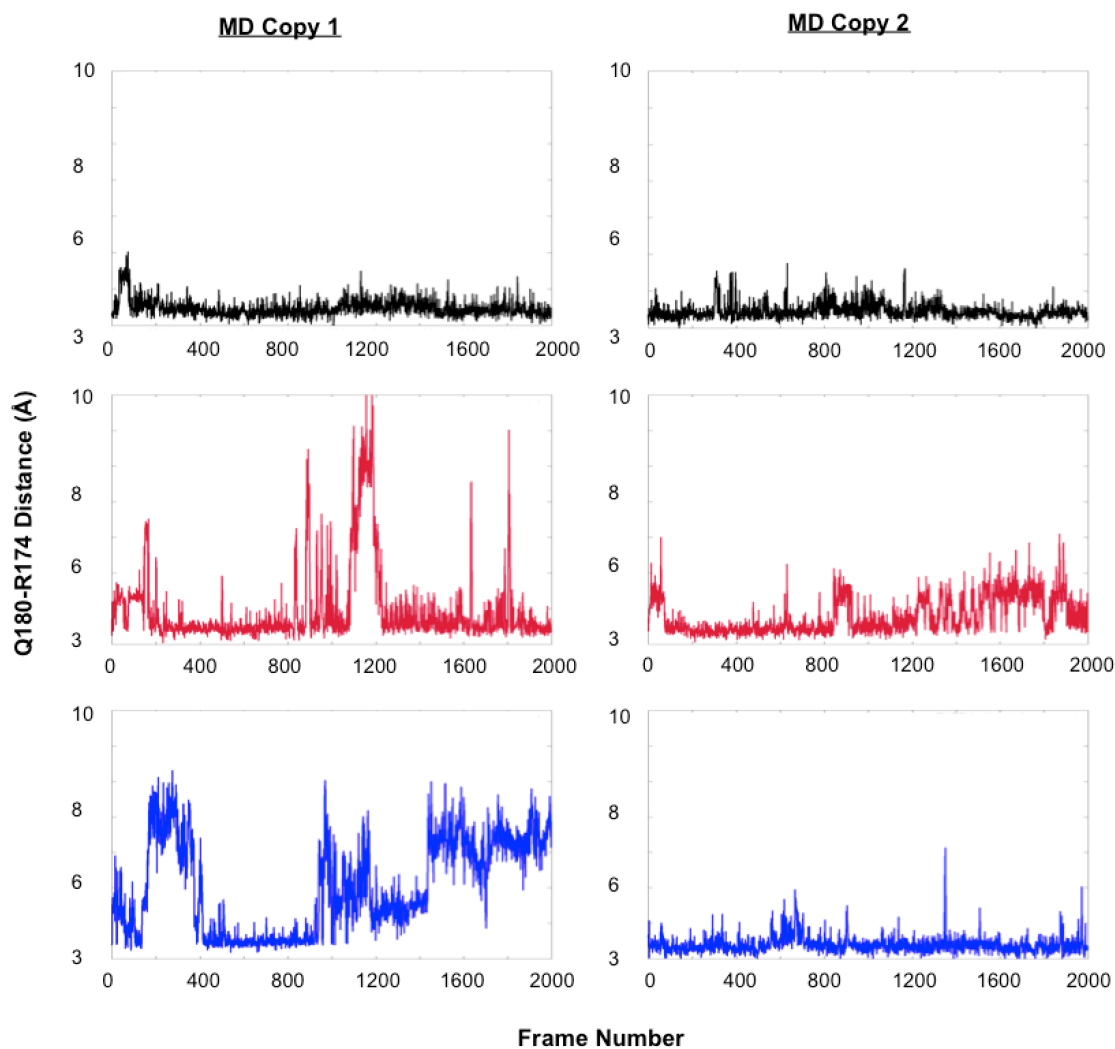


Figure 2.S2: Analysis of the Q180-R174 salt-bridge. The distance between the negatively- and positively-charged atoms in the Q180 and R174 residues are shown, where a stable salt-bridge is defined as 3.5Å. The salt-bridge in wildtype, R175H with zinc, and R175 without zinc are shown as a black, red, and blue line respectively for each individual M copy. Disruption of this salt bridge is seen in the R175H mutant systems.

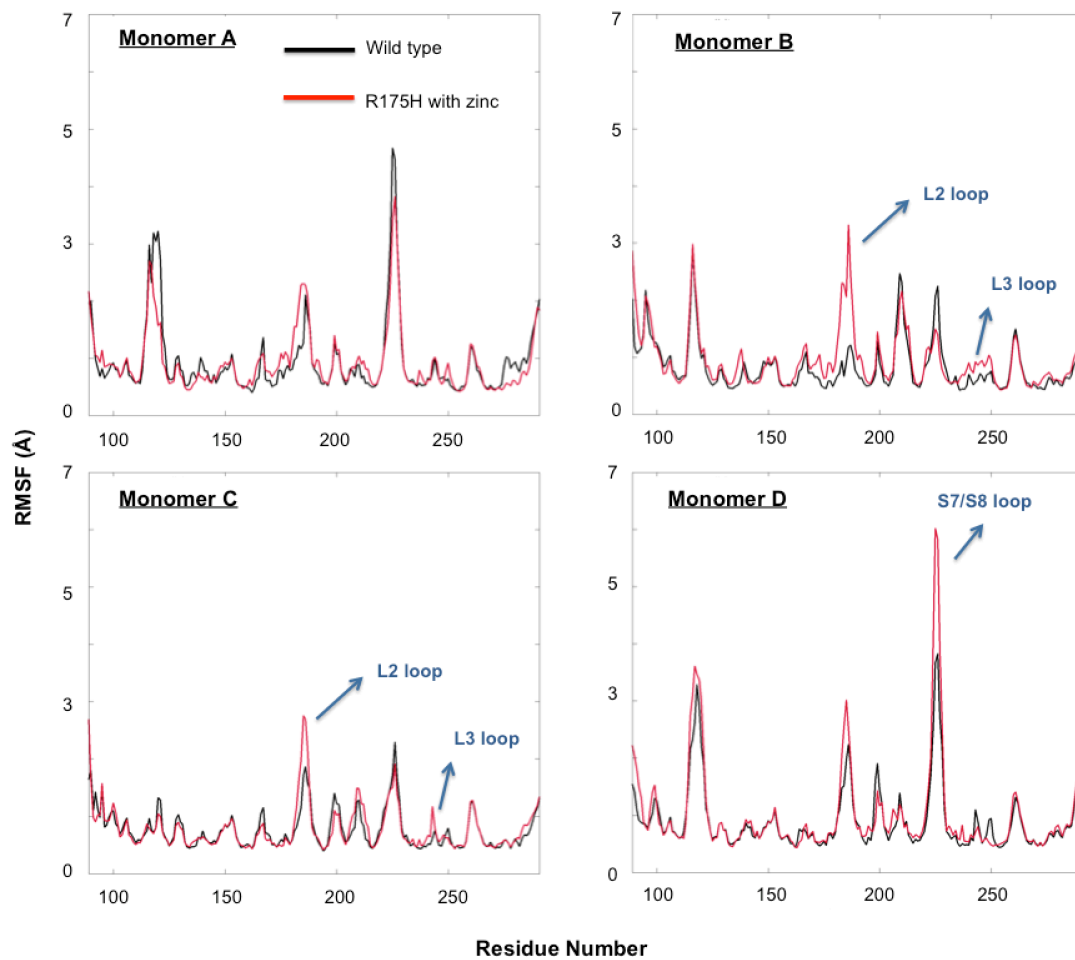


Figure 2.S3: Root-mean-square fluctuations of DBD: compare wildtype to R175H with zinc. The motifs of the DBD with increased flexibility in the R175H with zinc system are labeled.

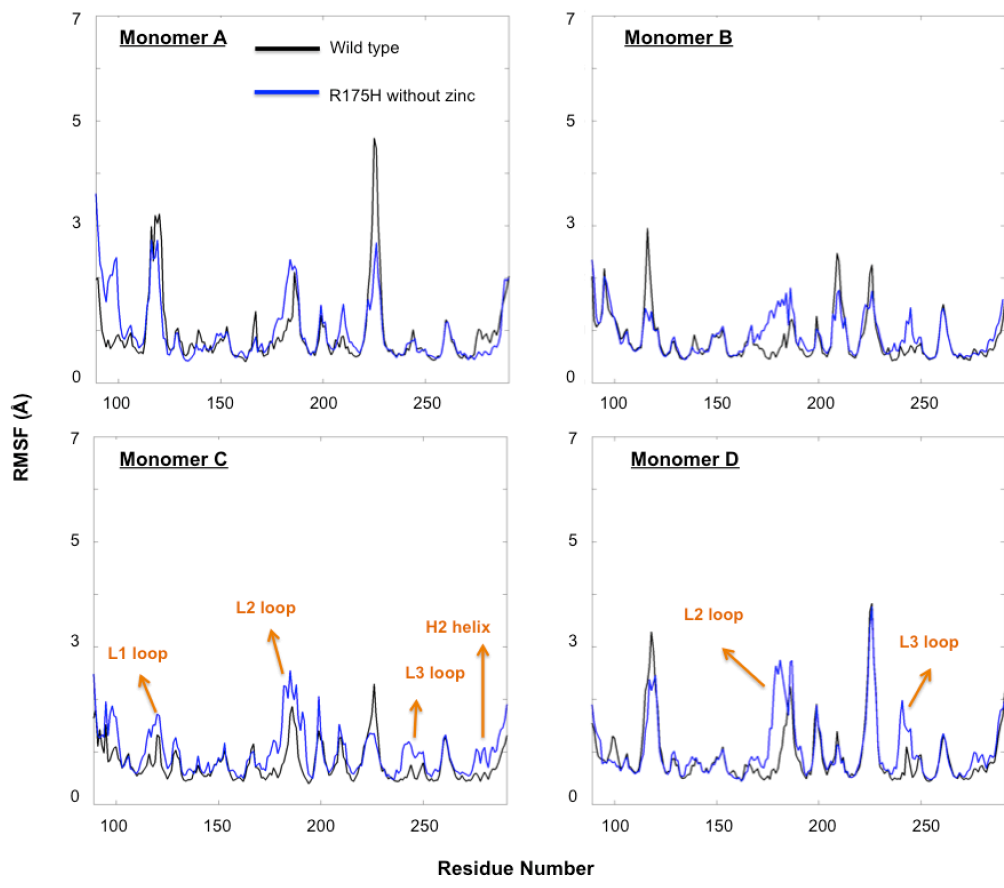


Figure 2.S4: Root-mean-square fluctuations of DBD: compare wildtype to R175H without zinc. The RMSF is shown for each monomer, and the motifs within the DBD where increased flexibility is seen for R175H without zinc system are labeled.

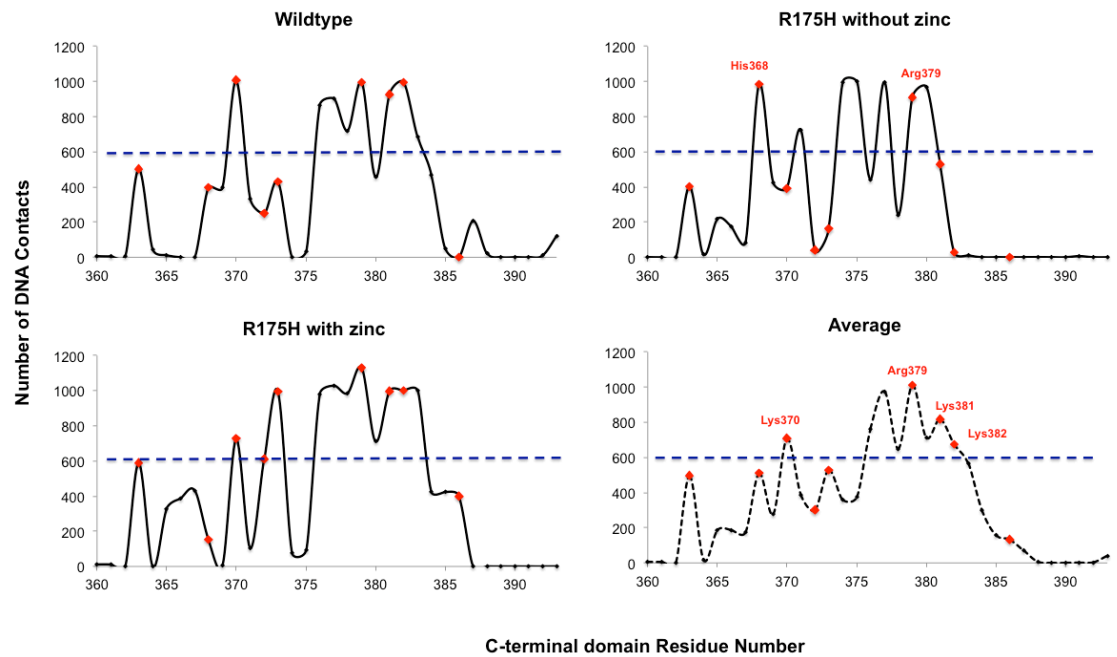


Figure 2.S5: CTD footprint analysis based on DNA contacts. The number of DNA contacts (CTD residues that come within 3.5Å of DNA) across all 4 monomers made for each CTD residue number is shown for each system. The average across all three p53 systems is also shown in the lower right panel. The positively charged residues in the CTD are highlighted in red. A value cutoff for the number of DNA contacts formed of 600 is selected (dashed blue line) since it is at least half of the maximum CTD-DNA contacts. The positively charged CTD residues that meet this cutoff are labeled.

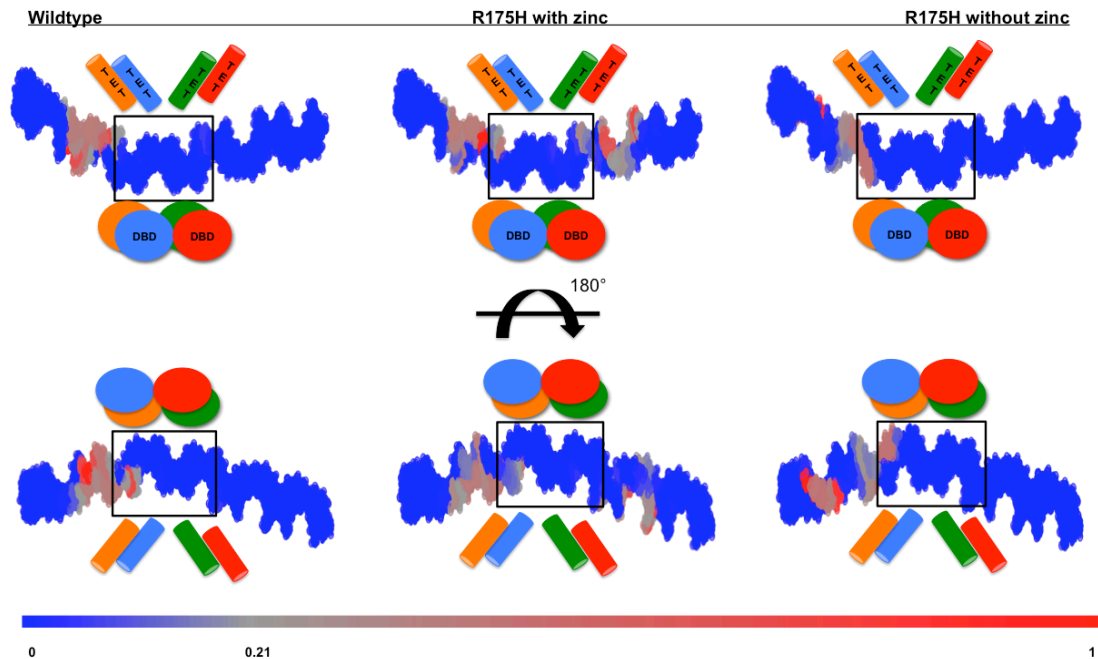


Figure 2.S6: DNA footprint analysis based on CTD contacts. The DNA bases that come within 3.5 Å of the CTD residues are mapped onto the DNA for each p53 system. The numbers of CTD contacts are normalized, ranging from 0 (no CTD contacts) to 1 (maximum number of CTD contacts). A cartoon of the DBD and TET domains are depicted to highlight the orientation of the DNA, and the DNA response element is highlighted with a black box.

Chapter 2, in part is currently being prepared for submission for publication of the material. Offutt, Tavina L.; Jeong, Pek U.; Demir, Özlem; Amaro, Rommie E. The dissertation author is the primary investigator and author of this paper.

Chapter 3

Enhancing Virtual Screening Performance of Protein Kinases with Molecular Dynamics

Simulations

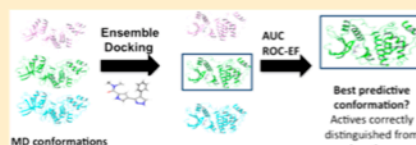
Enhancing Virtual Screening Performance of Protein Kinases with Molecular Dynamics Simulations

Tavina L. Offutt, Robert V. Swift, and Rommie E. Amaro*

Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92092-0340, United States

S Supporting Information

ABSTRACT: In silico virtual screening (VS) is a powerful hit identification technique used in drug discovery projects that aims to effectively distinguish true actives from inactive or decoy molecules. To better capture the dynamic behavior of protein drug targets, compound databases may be screened against an ensemble of protein conformations, which may be experimentally determined or generated computationally, i.e. via molecular dynamics (MD) simulations. Several studies have shown that conformations generated by MD are useful in identifying novel hit compounds, in part because structural rearrangements sampled during MD can provide novel targetable areas. However, it remains difficult to predict a priori when an MD conformation will outperform a VS against the crystal structure alone. Here, we assess whether MD conformations result in improved VS performance for six protein kinases. MD conformations are selected using three different methods, and their VS performances are compared to the corresponding crystal structures. Additionally, these conformations are used to train ensembles, and their VS performance is compared to the individual MD conformations and the corresponding crystal structures using receiver operating characteristic curve (ROC) metrics. We show that performing MD results in at least one conformation that offers better VS performance than the crystal structure, and that, while it is possible to train ensembles to outperform the crystal structure alone, the extent of this enhancement is target dependent. Lastly, we show that the optimal structural selection method is also target dependent and recommend optimizing virtual screens on a kinase-by-kinase basis to improve the likelihood of success.



■ INTRODUCTION

In drug discovery projects, high-throughput biochemical screens (HTS) are commonly used to identify pharmacologically active compounds. Despite extensive automation these screens still require expensive equipment and labor, contributing to the ~\$1.8 billion cost to bring a drug to market.^{1,2} Therefore, improving the efficacy of the hit discovery process has the potential to benefit multiple stakeholders, from patients to pharmaceutical companies. Structure-based virtual screening (SBVS) utilizes structural information from the drug target to predict ligand-protein interactions and can be more cost-effective than traditional HTS alone.³ During SBVS, ligand-protein interactions are used in a scoring function that predicts the binding affinities of a database of compounds against a drug target. These predicted affinities can then be used to prioritize a smaller subset of compounds for experimental testing.⁴ A good scoring function reliably distinguishes known active compounds from inactive compounds.

While it is common practice to use a receptor whose coordinates are determined by X-ray crystallography for VS, the approach has limitations. For example, a single crystal structure only captures one conformation and provides limited information about a protein's dynamic behavior, which can be an important regulator of ligand binding, as explained in two contemporary models. In the late 1950s, Koshland suggested

that ligand binding induces a conformational change in its cognate target that enhances ligand-binding affinity.⁵ With the advent of energy landscape theory, this concept was extended to the conformational selection method, which states that ligand binding biases conformational populations toward a single state.^{6–10} Consistently ignoring the importance of protein dynamics can have a detrimental impact on VS outcomes. This can occur when the crystallographic binding site conformation is not predictive, and large numbers of false positives and false negatives result. Therefore, it is important to consider the dynamic properties of proteins when predicting ligand-binding affinities. To address the importance of protein flexibility in SBVS, ensemble docking, which docks ligands into multiple target conformations, was developed.

There are several ways to generate protein conformations for ensemble docking. One can use experimentally determined protein structures solved using X-ray crystallography or NMR, in various ligand-bound and unbound states.^{11–20} However, the amount of time and expertise required to perform these experiments limits their utility. Alternatively, molecular dynamics (MD) simulations can reveal novel protein conformations with practical value to ensemble docking virtual

Received: May 6, 2016

Published: September 23, 2016

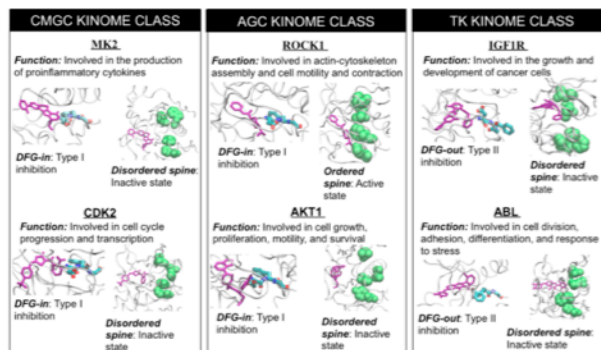


Figure 1. Protein kinases involved in study. For each protein kinase the crystallographic inhibitor is shown as a purple stick. The orientation of the DFG (Asp-Phe-Gly) motif was used to determine the inhibition type and is depicted as licorice colored by atom type (C cyan, O red, N blue). The assembly of the hydrophobic spine residues⁴⁸ (shown as green balls) was used to determine the conformational state of the protein kinase.

screens. A number of studies have successfully used MD-generated ensembles to identify active compounds.^{21–27} Nevertheless, despite efforts to determine how the use of MD structures affects chemical database enrichment,^{28–31} defining protocols for selecting MD structures across various protein targets for virtual screens remains difficult. Also, it is challenging to know, a priori, which protein targets will benefit from incorporating multiple target conformations in virtual screening experiments.

Here, we determine if MD-generated ensembles can be trained to maximize VS performance for six protein kinases. By analyzing the trained ensembles on an independent test set, we determine whether the addition of MD snapshots boosts the VS performance compared to the crystal structure alone. Specifically, we explore the impact of protein dynamics and ensemble training on VS performance for a set of kinase targets.

Protein kinases mediate most of the signal transduction in eukaryotic cells via phosphorylation of substrates.³² A comprehensive genomewide study found that there are ~500 protein kinases in humans, comprising ~1.7% of the human genome.³³ They are involved in many cellular processes including: metabolism, transcription, cell cycle progression, cytoskeletal rearrangement and cell movement, apoptosis, and differentiation.³⁴ Given the importance of phosphorylation, it is not surprising that abnormal phosphorylation can lead to cancers,^{35–37} cardiovascular diseases,³⁸ neurodegenerative diseases,³⁹ inflammatory diseases,^{40,41} and diabetes,⁴² thereby making protein kinases an important drug target. Consistently, an analysis of FDA-approved drugs since the 1980s indicated that kinases have surpassed GPCRs as the most sought-after targets for cancer treatments.⁴³ To date, the US Food and Drug Administration have approved 27 small molecule protein kinase inhibitors and 1 lipid kinase inhibitor.⁴⁴

Although drug discovery for protein kinases has achieved a great deal of success, several significant challenges remain in the development of future drugs. First, evolutionary pressure results in the accumulation of point mutations in the kinase domain, which compromises inhibitor potency and leads to long-term drug resistance.⁴⁵ Second, the conserved architecture of the kinase domain within a class of protein kinases (for example,

JAK kinases), makes obtaining selectivity challenging.^{40,46} This lack of specificity sometimes leads to adverse side effects. Third, the current kinase inhibitors on the market only cover a small subset of the human kinome, with 18 of the 27 approved covering only three out of more than 90 groups of tyrosine kinases, BCR-Abl, ErbBs, and VEGFRs. Given these shortcomings and the importance of the target, there is a need to improve kinase drug discovery—optimizing the enrichment of actives in virtual screening methods by using ensemble docking is one important avenue.

In the present study, we determine if MD structures result in enhanced virtual screening performance compared to the crystal structure alone. The effect of structural selection is examined by considering three methods: RMSD clustering, volume-based clustering and random selection. The impact of training heterogeneous ensembles that consist of both MD conformations and a crystal structure is also considered. Performance analysis is conducted using receiver operating characteristic (ROC) curve metrics, and the analysis is conducted for six protein kinases that cover three different kinase classes.

■ MATERIALS AND METHODS

Protein Kinase Systems. Six protein kinases from the directory of useful decoys-enhanced (DUD-E)⁴⁷ were included in this study. These include: (1) MK2 inactive conformation of MAP-kinase-activated protein kinase 2 (PDB code: 3M2W), (2) CDK2 inactive conformation of cyclin-dependent protein kinase 2 (PDB code: 4GCJ), (3) ROCK1 active conformation of rho-associated protein kinase 1 (PDB code: 2ETR), (4) AKT1 inactive conformation of serine/threonine-protein kinase AKT1 (PDB code 4GV1), (5) IGF1R inactive conformation of insulin-like growth factor 1-receptor (PDB code: 2OJ9), and (6) ABL inactive conformation of nonreceptor tyrosine kinase ABL1 (PDB code 2HZI). The assembly of the hydrophobic spine was used to determine the conformational state of each protein kinase (Figure 1).⁴⁸ Each protein kinase was aligned to the sequence of the cyclic adenosine monophosphate-dependent protein kinase (PDB 2CPK) to determine the residues that make up the hydrophobic spine. Visual analysis of the

hydrophobic spine assembly was conducted to determine the conformational state: an ordered hydrophobic spine indicates an active conformation, and a disordered spine reveals an inactive conformation.⁴⁸

The human kinome organizes all human kinases into 7 major groups, and the 6 kinases shown in Figure 1 cover three of these classes.³³ The CMGC kinase class, including MK2 and CDK2, consists of a diverse group of kinases named after cyclin-dependent kinases, mitogen-activated protein kinases, glycogen synthases, and CDK-like kinases. The AGC kinase class, including ROCK1 and AKT1, contains serine/threonine kinases regulated by cyclic AMP or lipids. The TK class, including IGF1R and ABL1, contains both receptor and cytosolic tyrosine kinases.

For all protein kinases, the crystal structures were chosen such that all activation loop residues were resolved with a resolution better than 3.0 Å (Table 1). For all protein kinases

Table 1. Protein Kinase Systems Setup for MD Simulations and VS Training

protein kinase	PDB code	atoms simulated ^a	total simulation time	actives ^b	decoys ^b
MK2	3m2w	54161	100 ns	50	3050
CDK2	4gcj	50644	100 ns	236	13688
ROCK1	2etr	74700	100 ns	49	3038
AKT1	4gv1	52854	100 ns	146	8030
IGF1R	2oj9	54844	100 ns	73	4526
ABL	2hzi	49818	100 ns	90	5220

^aThe total number of atoms in simulation includes the protein, ligand, ions, and explicit water molecules. ^bThe number of actives and decoys used in the trained models are shown. These numbers are the same across the training and test sets for both clustering methods.

except CDK2 and AKT1, the inhibitor-bound complex referenced in the DUD-E data set was used for the study, whose kinase conformation state matched the actives in the data set. In the case of CDK2 and AKT1, whose DUD-E crystal structure contained missing activation loop residues, alternative crystal structures that matched the conformational states of the DUD-E crystal structures were selected. Also, it should be noted that the ROCK1 kinase also included the N-terminal domain; for the remaining five kinases, the N-terminal domain was excluded, and only included the catalytic domain.

Preparation of Systems for MD Simulations. The six inhibitor-bound protein kinase crystal structures were obtained from the Protein Data Bank (PDB).⁴⁹ Using Schrödinger's Protein Preparation wizard version 2014-4, all six protein kinases were prepared.⁵⁰ Both ABL and ROCK1 kinases were crystallized as dimers; chain B was deleted and chain A was retained for both crystal structures. The remaining protein kinases were crystallized as monomers.

During the import and process step, the following boxes were checked for all six crystal structures: assign bond order, add hydrogens, create zero-order bonds to metals, create disulfide bonds, convert selenomethionines to methionines, delete waters beyond 5 Å from heteroatom groups, fill in missing side chains using Prime, and fill in missing loops using Prime.^{51–53} MK2 contained two missing residues in the N-terminus region and seven in a loop region. CDK2 only contained two missing N-terminus residues. For ROCK1, there were five N-terminus and ten C-terminus residues missing. For AKT1, there were two N-terminus residues, five residues in a

loop region, and three residues in the C-terminus missing. For IGF1R, there were four missing residues within the N-terminus and eight missing residues in a loop region. For ABL kinase, there were seven missing residues at the N-terminus. The missing N-terminus residues were ignored since they were located at the beginning and were far from the active site. However, the missing residues within a loop region were built in using Prime.

During the second preparation step, crystallization molecules and ions were deleted (MK2 magnesium ion; CDK2 four 1,2-ethanediol molecules; AKT1 four glycerol molecules; ROCK1, IGF1R, and ABL no crystallization molecules or ions present). AKT1 contained a phosphorylated threonine residue (Thr308-phospho), which was mutated back to threonine.

During the final preparation step, water molecules with less than three hydrogen bonds to protein residues were deleted. The protonation states of residues were assigned at pH 7 using PROPKA.^{54–56} Hydrogen bonds were optimized, followed by an all-atom minimization with termination based on convergence or reaching a heavy atom RMSD of 0.30 Å using the OPLS 2005 force field.⁵⁷ The resulting inhibitor–protein complexes were built for MD simulations in Amber14 xLeap.^{58,59} Antechamber was used to determine the atom types, bond orders, atomic partial charges, and assign force field parameters to the inhibitors using the gaff force field.^{60,61} The kinase–inhibitor complex was neutralized using chloride or sodium ions as described by Joung and Cheatham.⁶² The TIP4PEWBOX water model⁶³ was used to solvate the inhibitor 10 Å in the x-, y-, and z-direction.

MD Workflow. All-atom explicit-solvent MD simulations were performed for each inhibitor-bound protein kinase on GPUs using the CUDA version of pmemd in AMBER14.^{64,65} The general MD workflow consisted of three stages: minimization, equilibration, and production. The prepared systems were minimized in four steps using the steepest descent minimization method as follows: (i) minimization of the protons, while restraining the protein, ligand, and solvent; (ii) minimization of the solvent, while restraining the protein and ligand; (iii) minimization of the ligand and solvent, while restraining the protein; (iv) minimization of the protein side chains and water, while restraining the protein backbone; (v) minimization of all atoms in the system. Harmonic force constraint energy of 10 kcal/mol-Å² was used for the restrained minimizations. The systems were then equilibrated at 300 K and 1 atm for 200 ps with backbone restraints using the NPT ensemble. The backbone restraints were then removed, and the system was allowed to equilibrate for an additional 200 ps. MD was run for 5 ns in the NVT ensemble using the SHAKE algorithm,⁶⁶ and restart files were written every 1 ns. These restart files were used to begin five 20 ns NPT simulations at 300 K and 1 atm with a 2 fs time step. A total simulation time of 100 ns was generated for each protein kinase target.

Selection of Protein Conformations. After the MD simulations, the solvent and neutralizing ions were removed from each system. The 20 ns MD trajectories were loaded every 40 ps, resulting in 2500 MD snapshots, or frames, for each protein target. Residues within 10 Å of the inhibitor were selected and defined as the active site. MD trajectories were aligned on active site α carbon atoms using cpptraj.^{58,59} The aligned trajectories were clustered using two different clustering methods: (1) a Gromos RMSD-based method^{67,68} and (2) POVME 2.0.⁶⁹

Using the Gromos algorithm, pairwise root-mean-square deviations were calculated for all the active site heavy atoms for each frame. MD frames were clustered together if their RMSD value was below the specified cutoff. The cutoff value for each protein kinase system was selected using the following criteria: (i) there were no more than 40 clusters, (ii) 90% of the trajectory was within the first 10 clusters, and (iii) there were no more than 5 clusters with 1 single frame. The cluster centroids, or the representative for each cluster, of the number of clusters that contained at least 80% of the MD trajectory were selected to make up the ensemble for virtual screening.

The active site was also clustered based on its volume and shape using POVME 2.0. POVME 2.0 flooded the active site with equidistant grid points, and removed points that clashed with the receptor. The resulting grid points were used to calculate the active-site volume. These grid points were clustered, generating the same number of clusters as in the RMSD-based method, using their Tanimoto similarity scores. The cluster centroids were extracted for virtual screening.

Frames were also randomly selected from the MD trajectory in which the number of random frames matched the same number of RMSD and POVME cluster centroids. For example, for systems with five cluster centroids, every 500th frame was selected. It is important to note that frames are not extracted in a time-dependent manner since one long 100 ns trajectory was not run; instead five 20 ns trajectories were grouped together.

Principal Components Analysis. GROMACS⁷⁰ was used to determine the two most dominant modes of motion of the active site of the MD conformations using principal components analysis (PCA). The trajectories aligned on the active site heavy atoms were used to remove translation and rotation before generating PCA. The variances of the atomic coordinates were determined using the first MD frame as the reference; these variances were used to generate the covariance matrix A . Diagonalization of the covariance matrix was used to identify eigenvectors 1 and 2. In other words, these variances were projected onto 2D space corresponding to the first two principal components or the components with the greatest amount of variance. Calculation of the covariance matrix A was conducted as follows, yielding the eigenvalues, λ : $A\mu = \lambda\mu$. A plot of these eigenvalues along eigenvectors 1 and 2 was used to compare how the cluster centroids represented the structural changes throughout the MD trajectory.

Bio3D⁷¹ was used to determine the principal modes of motion for the MD conformations and crystal structures. A PDB search using the UniProt ID and a BLAST search in the Bio3D R package was used to find all available inhibitor-bound crystal structures for all six protein kinases. A consensus-binding site was found between all crystal structures and MD conformations, and the conserved binding site residues were used for protein alignment. The Cartesian coordinates of the aligned conserved binding site residues were the elements of the covariance matrix. Diagonalization of the covariance matrix was used to derive principal components 1 and 2, and calculation of the matrix resulted in the eigenvalues. A plot of these eigenvalues along eigenvectors 1 and 2 was used to compare the PC space of the MD sampled and available inhibitor-bound crystal structures.

Ensemble Docking. For all cluster centroids from the MD simulations, water molecules and ions were removed. Schrödinger's Maestro protein preparation wizard was used to determine the correct atom types of the cluster centroids.⁵⁰ The protonation states used during MD simulations were

retained for docking. The inhibitor was used to generate the receptor grid, using the default settings.

The DUD-E database of compounds for each protein kinase target was used for the virtual screen.⁴⁷ This database consists of experimentally determined actives and property-matched decoys. For every active compound, 50 decoys are selected to ensure physicochemical similarity and topologically dissimilarity to each active. The bond orders, stereochemistry, hydrogen atoms, and protonation states were generated for the actives and decoys using Schrödinger's LigPrep OPLS_2005 force field.⁷²

The prepared active and decoy molecules were docked into their respective crystal structure and ensemble of cluster centroids using Schrödinger's Glide single precision (SP) scoring function.^{73–75} The crystallographic pose of the inhibitor in the inhibitor-kinase complex was reproduced to validate the use of the Glide scoring function (Figure S1).

Measuring VS Performance. To ensure adequate VS performance, the ability to distinguish active and inactive, or decoy, molecules should be determined before performing a prospective virtual screen. While there are various metrics available to determine how well a virtual screen performs, this study focuses on two metrics: (1) the area under the receiver operating characteristic curve (AUC) and (2) ROC-enrichment factor (ROC-EF). Both the AUC and ROC-EF are determined from the ROC curve, which plots the true positive fraction (TPF) against the false positive fraction (FPF) at various threshold settings.^{6,77} The global classification ability of a virtual screen is given by the AUC, which is identical to the probability that an active will be ranked ahead of a decoy. However, the ability of a VS to enrich actives ahead of decoys early in the ranked list is often of more interest in drug discovery applications and enrichment factors (EF) are used to measure this ability. "Traditional" EF measures the ratio of actives in an early portion of the ranked list to the ratio of total actives in the database screened.⁷⁸ While this commonly used metric provides a useful indication of early enrichment, its maximum value depends on the ratio of decoy to active molecules in the database.⁷⁹ This can be problematic when making performance comparisons across targets where the decoy to active ratio varies, as is the case for each of the targets considered in this work. Therefore, the ROC-EF metric does not suffer from the same liability. It is determined by calculating the ratio of the TPF (at a given FPF) to the FPF, $TPF(FPF)/FPF$.⁷⁸ Random classification is indicated by a value of 1 and perfect separation of actives and decoys is indicated by a value of FPF^{-1} . We select a FPF of 0.001 in measuring the early chemical enrichment. However, it should be noted that an alternative FPF can be used, and to our knowledge there is no standard or generally accepted protocol for selecting a FPF.

Training Ensembles to Optimize VS Performance. A recently developed method, *EnsembleBuilder*, was used to train ensembles to maximize AUC or ROC-EF.⁸⁰ The docking results of each centroid and crystal structure were merged together and randomly split into a training and test set, maintaining the same active-to-decoy ratio (Table 1). Using the training set, all combinatorial possibilities at each ensemble size was constructed and either AUC or ROC-EF values were used to rank the performance of the resulting ensembles. For example, given two cluster centroids and the crystal structure, (labeled A, B, and xtal, respectively), there are seven possible ensembles: three of ensemble size one (A, B, or xtal), three of sizes two (AB, A and xtal, and B and xtal), and one of ensemble

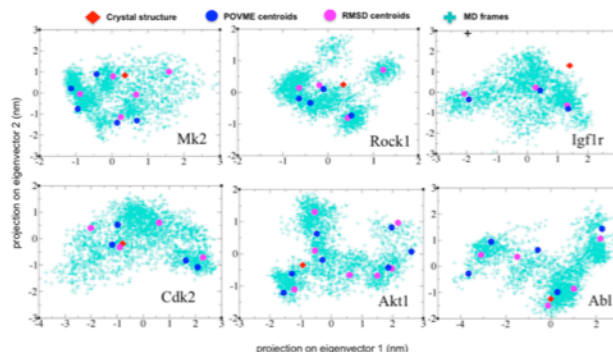


Figure 2. Comparison between how RMSD and POVME clusters the MD trajectory. Principal components 1 and 2 of the binding site in the MD snapshots, cluster representatives (centroids), and the crystal structure projected onto 2D space for each protein kinase are shown.

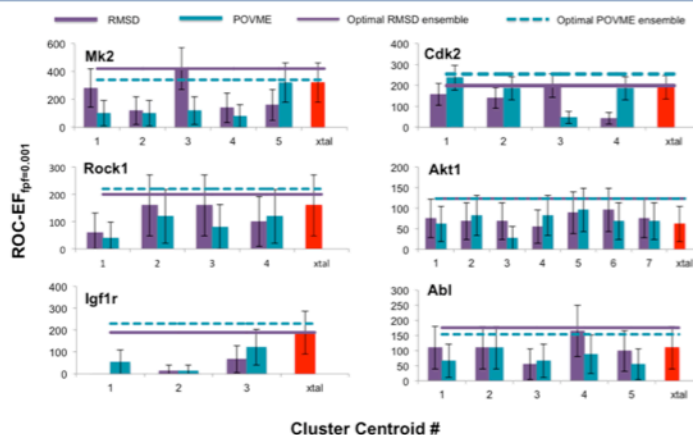


Figure 3. ROC-EF of the cluster centroids and crystal structures against the training set for each protein kinase. The ROC-EF of the optimal trained ensemble (the ensemble with the highest ROC-EF) using RMSD and POVME centroids reveal an increase in ROC-EF compared to the cluster centroids.

size three (A, B, and xtal). For each ensemble size, the best docking score value across all ensemble members was used to rank each compound,¹⁵ and AUC and ROC-EF values were determined from the resulting ranked lists. Finally, the ensemble combination with the largest AUC or ROC-EF was identified and retained. These best-performing ensemble combinations were used to screen the test set, and the resulting AUC and ROC-EF values were used to gauge prospective VS performance.

Quantifying VS Performance Gain. To compare the VS performance between the cluster centroids (or trained ensembles) and the crystal structure, the gain in AUC or ROC-EF was calculated as follows:

$$\text{AUCgain (\%)} = [\text{AUC}_{\text{centroid/ensemble}} - \text{AUC}_{\text{crystal}}] \times 100$$

$$\text{ROCEfgain (\%)} = (\text{ROCEf}_{\text{centroid/ensemble}} - \text{ROCEf}_{\text{crystal}}) \times 100/1000$$

To ensure the percent gain values are between 1 and 100, the ROC-EF is scaled by the FPF of interest.

RESULTS

We performed 100 ns of MD on each protein kinase and wrote out snapshots every 40 ps, which resulted in a total of 2500 frames. Docking into every MD frame is computationally expensive, and because of long conformational relaxation times, likely unnecessary. Therefore, various metrics are available to reduce the MD ensemble in a meaningful way without losing critical structural information about the dynamic behavior of

the protein. In this study, we utilized three methods for reducing the MD trajectories: RMSD and POVME clustering in addition to random selection. The resulting cluster centroids and random frames were used for VS experiments. The VS performances of the individual cluster centroids were compared to the crystal structure and random frames performance. Lastly, the performance of ensembles trained to maximize either the AUC or ROC-EF was also considered.

Comparison Between Structural Selection Methods. *Comparison between RMSD and POVME Clustering Methods.* Both RMSD and POVME clustering adequately capture large-scale conformational changes of the binding pocket that occur during the MD simulations (Figure 2). Each method samples a collection of conformations that collectively represent the conformational space spanned by the first two principal components. However, the conformational representation used by each (binding pocket RMSD and shape) is different and results in a unique pool of cluster centroids for each method. To determine how these differences impact VS performance, we compared the VS performance of each cluster centroid to the corresponding crystal structure using the training set.

When considering global classification ability, as measured by the AUC, POVME yields the highest individual performing conformation (Figure S2), but this result is not consistent across all targets; both RMSD and POVME clustering return centroids with VS utility. For instance, RMSD and POVME both yield high performing MK2 centroids that perform identically (AUC = 0.97) and marginally outperform the crystal structure (AUC = 0.96), representing a 1% performance gain. This small gain is not surprising: the large crystal structure AUC value leaves little margin for improvement, and any performance gain must necessarily be small. Also while performance gains varied across targets, high crystal structure AUC values and small performance gains were the norm. For instance, while AKT1 and ABL realize the largest AUC gains using the highest performing conformer, these gains were only 5.53% and 5.51%, respectively. Similarly, the highest performing conformations of CDK2, ROCK1, and IGF1R resulted in AUC gains of 3.62%, 3.67%, and 2.83%, respectively. While baseline crystal structure AUC values were high across all targets, this was not the case for early enrichment, where crystal structure performance was less than perfect, a case we consider next.

Similar to the global classification results, when early enrichment is considered, as measured by ROC-EF values, the clustering method that identifies the highest performing conformation is target dependent, and both methods reveal centroids with VS utility (Figure 3). For instance RMSD produces the highest performing cluster centroids for MK2, ROCK1, and ABL, while POVME yields the best performers for CDK2 and IGF1R. For AKT1, both clustering methods yield a conformer with the same ROC-EF of 95.89. Additionally, for four out of six targets (MK2, CDK2, AKT1, and ABL), both methods were able to identify conformations that outperformed the crystal structure. At 10%, the largest gain in early enrichment ability was realized by an MK2 centroid, which was identified by RMSD clustering. For CDK2, AKT1, and ABL, gains using the highest performing conformer are 4.66%, 3.42%, and 5.49%, respectively. In contrast to these performance gains, neither ROCK1 nor IGF1R yielded centroids that outperformed the crystal structure. However, the highest performing ROCK1 centroid performed identically

to the crystal structure (Figure 3). For IGF1R, where all centroids perform worse than the crystal structure, we note that the PCA plots of the crystal relative to the MD frames indicates little overlap, which may explain this ROC-EF difference.

Comparison between Clustering Methods and Random Selection. The randomly selected frames sample conformations that collectively represent the conformational space spanned by the first two principal components for AKT1 (Figure S3). However, for the remaining five protein kinases, the random frames samples a subset of PC space. Similarly to the clustering methods, the conformational representations of the random frames are unique and differ from both RMSD and POVME cluster centroids. To determine if clustering the MD simulation in a meaningful way versus randomly selecting frames has an impact on the VS performance, the VS performance of the random frames were compared to the cluster centroids against the entire data set.

When considering global VS performance, the AUC of the random frames and cluster centroids are generally comparable, as indicated by the overlapping confidence intervals (Figure S4). For two systems, ROCK1 and IGF1R, the randomly selected frames yield the single highest MD conformer. However, these AUC values are not statistically higher than the single highest yielding cluster centroids. Therefore, we conclude that the structural selection method does not impact global VS performance. Next, we considered the early enrichment metric in comparing the structural selection methods.

Similar to global VS performance, when early enrichment is considered, the randomly selected frames yield similar ROC-EF values to the cluster centroids for majority of the targets, with the exception of AKT1 and IGF1R (Figure S5). For AKT, there are four random frames that have a higher EF than all cluster centroids; however, this increase is not statistically significant. For IGF1R, the single best MD conformer comes from the random selection method, which is not statistically higher than the single highest cluster centroid. Similarly to the results seen with AUC, the structural selection method does not impact the early enrichment of VS.

Even though our results show that the use of clustering methods does not boost VS performance over random selection, the clustering methods does a better job in representing the MD conformations as shown with PCA. Therefore, we limit the use of randomly selected frames here, and only consider the cluster centroids for ensemble training. To determine whether synergism between cluster centroids and the crystal structure could result in performance that exceeded that of any individual conformation, we used these structures to train ensembles.

Comparing Trained Ensemble and Crystal Structure VS Performance. Training ensembles on AUC yielded at least one ensemble with a higher AUC than the crystal structure against the training set (Figure S6). The optimal trained ensembles contained no more than four members across all six targets (Table 2). For MK2 and CDK2, the members consist of RMSD centroids, and POVME centroids for the remaining targets. Interestingly, the optimal trained ensemble's AUC value is statistically significantly higher than the crystal structure across all six targets ($p < 0.05$). The largest AUC gain using the optimal ensemble over the crystal is seen with ABL with a gain of 8.62%, while the smallest AUC gain (1.71%) is seen with MK2. For CDK2, ROCK1, AKT1, and IGF1R, the AUC gains are 5.20%, 6.93%, 7.06%, and 3.09% respectively. Next, we

Table 2. Global Performance of the Virtual Screen (AUC) of the Optimal Trained Ensemble against the Test Set

kinase	ensemble size	cluster method	ensemble members	higher than xtal?	statistically significant ^a
MK2	3	RMSD	centroids 1, 3, and xtal	yes	yes
CDK2	3	RMSD	centroids 2, 3, and 4	yes	yes
ROCK1	3	POVME	centroids 2, 3, and xtal	yes	yes
AKT1	4	POVME	centroids 4, 5, 6, and xtal	yes	yes
IGF1R	2	POVME	centroids 2 and 3	yes	yes
ABL	3	POVME	centroids 2, 3, and 4	yes	yes

^aStatistical significance is determined at the 95% confidence level; $p < 0.05$. ^bThe centroids highlighted in bold contributed the most to the VS performance.

determined if training ensembles on ROC-EF would result in increased early enrichment performance over that of the crystal structure.

Similarly to training on AUC, ensembles trained on ROC-EF resulted in at least one ensemble that outperformed the crystal using the training set (Figure 4). All optimal trained ensembles contained no more than three members (Table 3). The members consist of RMSD centroids for MK2 and ABL, and POVME centroids for CDK2, ROCK1, and IGF1R.

For AKT1, there are two optimal trained ensembles (RMSD size 3 and POVME size 2) as they have the same ROC-EF value. The optimal trained ensembles' ROC-EF is statistically significantly higher than the crystal for CDK2 and AKT1 ($p < 0.05$), with a gain of 6.36% and 6.16% respectively. For MK2, ROCK1, IGF1R, and ABL, the ROC-EF gains are 10.00%, 6.00%, 4.05%, and 6.59%, although these gains are not statistically significant ($p > 0.05$). Interestingly, results demonstrate that trained ensembles of cluster centroids and

Table 3. Early Chemical Enrichment of Actives (ROC-EF_(pf=0.001)) of the Optimal Trained Ensemble against the Test Set

kinase	ensemble size	cluster method	ensemble members ^a	higher than xtal?	Statistically significant ^b
MK2	1	RMSD	centroid 3	yes	yes
CDK2	2	POVME	centroids 1 and 4	yes	no
ROCK1	2	POVME	centroid 2 and xtal	no	no
AKT1 ^b	2	POVME	centroids 4 and 5	yes	no
	3	RMSD	centroids 5, 6, and 7	yes	no
IGF1R	3	POVME	centroids 2, 3, and xtal	yes	no
ABL	2	RMSD	centroids 2 and 4	yes	yes

^aStatistical significance is determined at the 95% confidence level; $p < 0.05$. ^bTwo ensemble sizes are shown because both result in the same ROC-EF value which is the highest. ^cThe centroids highlighted in bold contributed the most to the VS performance.

the corresponding crystal structure yielded VS performance that exceeds that of the crystal structure. To determine if the optimal trained ensembles performance was due to synergism between structures or due mostly in part to a single high performing conformer; we also compared the optimal trained ensemble's VS performance to the single highest performing MD conformer.

VS Performance of Trained Ensembles vs Cluster Centroids. Ensembles trained on AUC resulted in at least one ensemble with a higher AUC than the cluster centroids against the training set (Figure S2). However, this increase in AUC using the optimal trained ensemble over the single highest performing conformer is not statistically significant for any targets ($p > 0.05$), with gains of only 0.85%, 1.58%, 3.26%, 1.53%, 0.26%, and 3.11% for MK2, CDK2, ROCK1, AKT1,

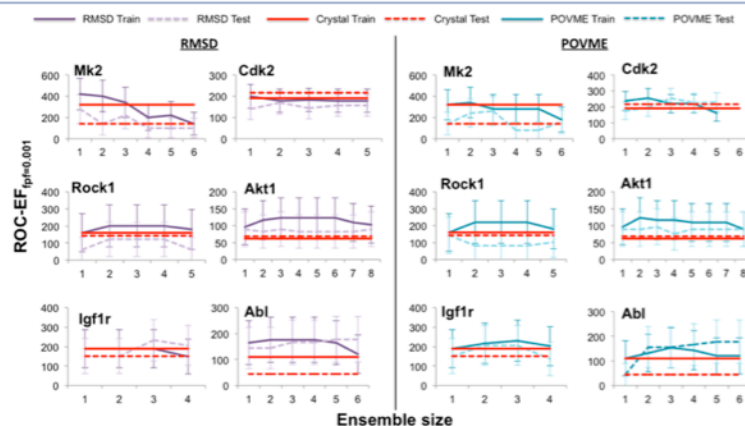


Figure 4. Trained ensemble sizes and crystal structures ROC-EF ($fpf = 0.001$) values against the training and test set across all six protein kinases. The 95% confidence intervals overlap between the training and test set, validating the training method.

IGF1R, and ABL respectively. However, it is noteworthy that the highest individual performing conformer is a member of the optimal trained ensemble for all targets except CDK2 (Table 2). Recall for CDK2, POVME yields the single highest performing conformation; however, the optimal trained ensemble consists of RMSD cluster centroids. Therefore, if we look within the RMSD cluster centroids only, we see that the optimal ensemble contains the highest performing RMSD conformer, thereby supporting the results seen with the other five targets. In addition to being a member of the optimal ensemble, the single highest performing conformation contributes the most to the VS performance of the optimal trained ensemble for all targets except AKT1 (Table 2, centroids highlighted in bold). This means that the single highest performing conformer was more successful at correctly ranking actives ahead of decoys than any other ensemble member. While results suggested that the global performance of the trained ensembles is due to the single highest performing centroid, next we investigated if the same was true for early enrichment.

Ensembles trained to maximize ROC-EF resulted in at least one ensemble with a higher ROC-EF than the cluster centroids against the training set for all targets except MK2 (Figure 3). This exception is due to the fact that the single highest performing conformation yields a higher ROC-EF than all trained ensembles. A statistically significant ($p > 0.05$) higher gain in ROC-EF using the optimal trained ensemble over the single highest performing conformation is seen for IGF1R only, with a gain of 10.81%. For the remaining four targets, CDK2, ROCK1, AKT1, and ABL, the gains in ROC-EF are 1.69%, 6.00%, 2.74%, and 1.10% respectively, although this enhancement is not statistically significant ($p > 0.05$). Similarly to the members of the optimal ensemble trained on AUC, the members of the optimal trained ensembles on ROC-EF contain the single highest performing conformation across all six targets (Table 3). Also, the single highest performing conformation is the largest contributor to the optimal ensemble for all kinases except CDK2, suggesting the early enrichment performance of the optimal trained ensembles is due to the single highest performing conformer (Table 3, centroids highlighted in bold). Next we measured the VS performance of the trained ensembles against a test set of compounds it has never seen before in order to predict how the trained ensembles would perform prospectively.

Performance of Trained Ensembles Against the Test Sets. The VS performance of the trained ensembles against the test set validates the ensemble training method (Figures 4 and S6). Since the AUC and ROC-EF values on the test set are within the confidence intervals of the values against the training set, this suggests that the trained models may perform similarly prospectively. Similar to the analysis against the training set, we compared the VS performance of the optimal trained ensembles' and crystal structure against the test set.

The optimal trained ensembles resulted in slightly higher AUC values than the crystal structures against the test set across all six protein kinases (Table 2). This enhancement in AUC is statistically significant across all six targets ($p < 0.05$), similar to the optimal ensemble's performance against the training set. The AUC gains are 2.16%, 7.26%, 2.61, 5.73, 6.03%, and 3.38% for MK2, CDK2, ROCK1, AKT1, IGF1R, and ABL respectively; these gains are modest as expected due to the large crystal AUC values. Next, we compared the ROC-

EF of the optimal trained ensembles' and crystal against the test set.

The optimal trained ensemble resulted in a higher ROC-EF than the crystal structure against the test set for all targets except ROCK1 (Table 3). The enhancement in ROC-EF is statistically significant for MK2 and ABL only ($p < 0.05$), differing from the results against the training set where the ROC-EF is statistically significantly higher for CDK2 and AKT1. Similar to the early enrichment performance against the training set, the largest ROC-EF gain is seen for MK2 with a gain of 14%. Differing from the performance against the training set, no ROC-EF gain is seen for ROCK1, and the optimal trained ensemble and crystal perform identically (Figure 4). The gains in ROC-EF for CDK2, AKT1, IGF1R, and ABL are 3.81%, 2.74%, 8.22%, and 13.33%, respectively. We also compared the VS performance of the optimal trained ensembles to the single highest performing conformer against the test set. The AUC and ROC-EF of the optimal trained ensemble was either the same or higher than the single highest performer against the test set (Figures S7 and S8). However, this increase is not statistically significant ($p > 0.05$) for any target, so we limit discussion of our results here.

DISCUSSION

All Structural Selection Methods Yield Conformers that Enhance VS Performance. The impact that RMSD and POVME clustering methods have on VS performance is comparable. Both clustering methods selected at least one conformer that performed as well or better than the crystal structure, as judged by both AUC and ROC-EF values. While POVME identified the single conformer with the highest AUC across all targets, the improvement was not statistically significantly higher than the AUC of the best performing RMSD conformer. Therefore, we cannot conclude that POVME identifies conformers that result in meaningful performance gains relative to RMSD clustering. Either clustering method can successfully cluster the MD trajectory and reveal conformers that will be successful in VS experiments.

While the impact that clustering and randomly selecting MD frames have on VS performance is comparable, there are differences with how the methods represent the MD conformational space. The use of clustering methods ensures that the large-scale conformational changes of the binding site are captured in the VS experiments, which held true for all six targets. However, randomly selecting frames may or may not capture these conformational changes as seen with majority of the protein kinases within this study. Since, there is no way of knowing a priori the correlation between VS performance and binding site conformation, it is important to represent all conformations sampled during MD in VS methods. Therefore, we still conclude that clustering on some physical property of the binding pocket is important in ensuring that all conformations sampled during MD are represented in VS experiments. In choosing which clustering method to utilize, factors other than VS performance comparison should be considered.

In order to reduce computational cost and time, it is optimal to choose one clustering method; this decision may be made based on the advantages and limitations with each method. RMSD-based clustering is commonly used and has proven successful in revealing conformations that enriched novel active compounds.^{21,27,81} However, degeneracy may be a major limitation. Two conformers with the same RMSD would be

grouped into the same cluster, even though visual inspection may reveal key structural differences. To overcome this limitation, clustering on the volume and shape of the active site may be an attractive alternative. Also, it may be easier to visualize differences in active site shapes and volumes as opposed to changes in side chain positions.

Training Ensembles Enhances VS Performance. Training ensembles improved the early enrichment of actives. Although we saw a statistically significant increase in AUC with the trained ensembles, it is important to highlight that the crystal structures yielded high AUC values (>0.9 for one kinase, >0.8 for three kinases). Therefore, there was little room for improvement in the global VS performance. Based on the ROC-EF values of the crystal structure, which were further from the maximum value, there was greater potential to realize improved early enrichment.

While it is important to optimize the overall performance of the virtual screen (AUC), very often in drug discovery projects the ability of the VS method to enrich actives early in the ranking lists is an important feature. This is especially critical in cases where thousands or even millions of compounds are screened in silico, and only a small subset of compounds can be experimentally tested (i.e., in academic laboratories or limited resource settings, or for targets that do not yet have their assays ported to high throughput frameworks). In providing a top fraction of compounds for experimental testing, it is crucial to reduce the number of recommended false positives and enhance the hit rate. A boost in this early enrichment of actives is seen when ensembles are trained on ROC-EF. However, in the examples presented here, this performance boost is only statistically significant for MK2 and ABL, suggesting that the ROC-EF values of the trained ensembles are comparable to the crystal structure for the remaining four protein kinases at a false positive fraction of 0.001. At this false positive fraction, the total numbers of compounds enriched in the early ranking lists were 18, 20, 8, 22, 19, and 23 for MK2, CDK2, ROCK2, AKT1, IGF1R, and ABL kinases. To the best of our knowledge, there does not appear to be a general protocol established for choosing false positive fractions with early enrichment. Therefore, it may make sense to choose a false positive fraction based on the available resources for experimental testing. For example, if in future prospective studies we are able to screen 80–100 compounds, we can select an optimal ensemble that gives a higher ROC-EF at a later false positive fraction, 0.01 for example. If we look at the ROC-EF for CDK2, ROCK1, IGF1R, and AKT at a later false positive fraction of 0.01, there are trained ensembles with a statistically significantly higher ROC-EF than the crystal structure (Figures S9 and S10).

MD Simulations Reveal at Least One Better Predictive Conformation than the Crystal Structure. One hypothesis for why trained ensembles enhance the virtual screening performance over the crystal structure is because the MD simulation finds at least one single conformation that is more predictive than the crystal structure. Consistently, the optimal trained ensembles contained the single highest performing centroid for majority of the targets. It is interesting that the conformer that contributed the most to the optimal trained ensemble's performance is never the crystal structure and always an MD centroid (Tables 2 and 3). These results align with the studies of Barril et al, who observed that ensemble member's that performed well alone also did well in ensemble docking.¹⁶ Our results suggest that performing MD simulations

on protein kinases may be worthwhile in virtual screens as one single MD conformation may significantly enhance virtual screening performance. While identifying a single high-performing MD conformation may be promising, identifying it requires docking to every MD snapshot or cluster centroid, which is limiting. It would be more straightforward to select the highest performing centroid a priori; however, there is little consensus about the target descriptors that would allow such a conformation to be identified. In an effort to address this, we explored whether there was a correlation between VS performance and the binding site volume, and found no correlation. Ellingson et al. investigated whether there was a correlation between several physicochemical and thermodynamics properties and descriptors of MD snapshots and enrichment factors.⁸² Some of these properties included the number of MD snapshots that make up the cluster, the trajectory frame number of the largest neighbor conformation that is used as the representative conformation, largest pairwise RMSD within the cluster, and the largest RMSD within the cluster to the representative conformation. Additional properties included the number of contact atoms in the binding site, propensity for ligand binding, number of hydrophobic contact atoms in the binding site, number of side chain contact atoms in the binding site, the number of contact residues in the binding site, the van der Waals (vdW) surface area, hydrophilic surface area, hydrophobic surface area, and vdW volume. However, this attempt to identify a meaningful feature set that could reliably recognize high performing MD snapshots was unsuccessful. Future studies that elucidate MD snapshot selection for enhanced VS performance will be groundbreaking and greatly enable the use of MD-generated ensemble docking approaches.

While MD simulations within this study revealed a single high performing conformation, NMR and X-ray crystal structures may have also reveal a single high performing conformer. Although, we did not explore the use of NMR or crystallographic ensembles, we speculate that the virtual screening performance would be similar to the results seen using MD-generated ensembles for different reasons.

NMR experiments provide an ensemble of protein conformations, where each conformer could be extracted for VS experiments. However, we suggest that training an ensemble of conformations obtained from NMR would yield similar VS performance as the MD-generated ensembles since both methods explore protein dynamics. Damm and Carlson make a similar conclusion in a study where they found that both MD and NMR captured similar structural variations of HIV-1 protease as indicated in their similar pharmacophore models.⁸³

While multiple crystallographic structures are not always available for proteins, this is not the case for protein kinases, with over 300 crystal structures for Cdk2 for example (Supporting Table S1). Therefore, all the available crystallographic structures for each protein kinase could be used to generate an ensemble to train for VS performance. It is in fact possible that the VS performance of crystal structure ensembles may differ from the MD-generated ensembles. However, since PCA analysis reveals an overlap between the binding site of the crystal structures and the MD conformations (Figure S11), we conjecture that the ensemble of crystal structures would perform similarly to the MD ensemble and limit our studies to MD-generated ensembles.

Smaller Ensemble Sizes Result in Optimal VS Performance. When looking at the performance values of

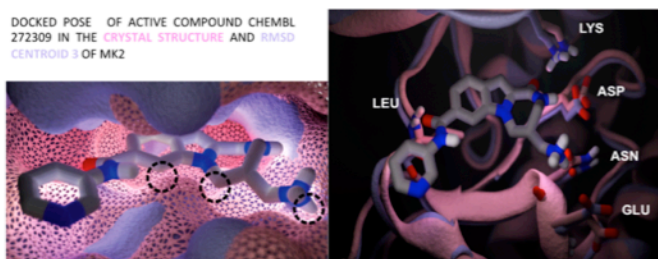


Figure 5. Docked pose of active compound, CHEMBL272309, reveals favorable interactions with RMSD centroid 3 (purple) and steric clashes (circled in left figure) with the crystal structure (pink). The compound makes H-bonding interactions with active site residues in the RMSD centroid 3, but interactions with the Glu, Asp, and Asn residues are lost when bound to the crystal structure. CHEMBL272309 is shown as licorice and is colored by atom type (C gray, N blue, O red, S yellow, H white).

ensembles trained to maximize either the AUC or ROC-EF, it is interesting to note that the optimal ensemble size is fairly small. For some protein kinases, adding more conformations either left the performance unaltered or degraded it. For example, if we look at the results of training AKT1 ensembles to maximize ROC-EF values (see Figure 4), the RMSD optimal ensemble was of size 3. For ensemble sizes 4, 5, and 6, the ROC-EF is the same as ensemble size 3; adding 3 additional MD centroids did not alter the early enrichment of actives. Furthermore, for ensemble sizes 7 and 8, the ROC-EF value decreases. Previous studies have shown similar results—the addition of more conformers can degrade performance.^{12,14–16,84–86} While ensemble docking studies, ours included, consistently highlight the importance of incorporating multiple protein conformations in VS experiments, that performance does not necessarily scale with ensemble size. Clearly, even though protein kinases may adopt multiple conformations, it appears that only a small number of conformations (1–4) are important for ligand binding, at least in the context of ensemble docking. The key to fully leveraging the power of ensemble docking will be developing strategies and protocols to find these crucial conformations, which remains an outstanding challenge to the field.

Structural Comparison between MD Conformer and Crystal Structure for MK2. To gain insight into why a MD cluster centroid may result in greater enrichment than the crystal structure, we analyzed key structural differences in an exemplar case. We determined several key structural differences that exist between the third RMSD centroid and the crystal structure of MK2. First, when we compare the actives enriched in the ranking lists at a false positive fraction of 0.001, we see that RMSD centroid 3 enriches an additional eight classes of actives that are not enriched by the crystal structure. One of these actives, CHEMBL272309, is unable to fit in the active site of the crystal structure, indicated by the steric clashes made with the crystal structure (Figure 5). Interactions between the active compound and the hinge region (Leu) and a C-lobe beta strand (Lys) are seen when bound to both the MD conformation and the crystal structure. However, three hydrogen-bonding interactions are lost with residues Asp, Glu, and Asn when bound to the crystal structure, which compounds the steric clash and greatly lowers the predicted binding affinity. Although the docked pose is shown for one active, the steric clash and suboptimal-hydrogen-bonding

network observed with the crystal structure is consistently observed for all of the actives uniquely enriched by the third RMSD centroid. These side-chain movements help explain the difference in early enrichment between the MD conformer and the crystal structure, and the subtle rearrangement underscores the difficulty of identifying a general set of structural features that reliably predicts conformations with strong classification ability.

CONCLUSIONS

In this study, we analyzed the VS performance of MD structures against six protein kinases in three different kinome classes. We compared two clustering methods to determine whether clustering the RMSD values of active site heavy atoms resulted in a significant advantage over clustering the active-site shape. We compared these clustering methods to random selection of MD conformations. We also determined if ensembles of MD structures and crystal structures could be trained to optimize AUC and ROC-EF, and we analyzed their performance on a test set of compounds.

Results from this study suggest that running MD is worthwhile in conducting virtual screens against protein kinases, because it may result in that at least one conformation is more predictive than the crystal structure. Further, ensembles trained to maximize the AUC or ROC-EF can result in better performance than using the crystal structure alone. However, the extent of this enhancement is system dependent. For the majority of the protein kinases in this study, it does not seem to matter whether MD structures are selected using RMSD or POVME clustering. We find that the virtual screening performance differs between targets. This held true for members of the same protein class and for members of the same kinome class. The performance variability across targets implies that optimizing virtual screening protocols on a target-by-target basis is a reliable way to improve the likelihood of a successful prospective virtual screen.

Although the results presented are encouraging, they are limited to the six protein kinases within this study. Exploring larger data sets will ultimately lead to a greater understanding of the fundamental promise and limitations of applying ensemble docking to kinase drug discovery. Parallel studies within our group are focusing similar analysis on nuclear hormone receptors. Consistent with our conclusions here, the VS performance within the same nuclear hormone receptor class

differs between each target, which supports the case for target-specific optimization prior to applying a VS method prospectively.

■ ASSOCIATED CONTENT

3 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00261.

Additional data (as described in the text) provided in supporting Figures S1–S11 and Table S1 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ramaro@ucsd.edu.

Author Contributions

T.L.O. performed the calculations and analysis and wrote the manuscript. R.V.S. developed the method, *EnsembleBuilder*, and assisted with manuscript preparation. R.E.A. directed the project and wrote the manuscript. All authors have given approval for the final version of the manuscript.

Notes

The authors declare the following competing financial interest(s): R.E.A. is a co-founder of Actavalon, Inc.

■ ACKNOWLEDGMENTS

The authors thank Susan Taylor, Alexandr Kornev, and Victoria Feher for useful discussions and guidance about protein kinases, Jamie Schiffer and Jeff Wagner for review of the manuscript, Jacob Durrant for assistance with manuscript preparation, Siti Jusoh for useful discussions about the work, and the Barry Grant lab, especially Xin-Qiu Yao, for help and assistance using the Bio3D package. This work was funded in part by the Director's New Innovator Award Program NIH DP2 OD007237 to R.E.A. Funding and support from the National Biomedical Computation Resource (NBCR) is provided through NIH P41 GM103426. T.L.O. is also sponsored by a San Diego Fellowship from UC San Diego.

■ ABBREVIATIONS

SBVS, structure-based virtual screening; VS, virtual screening; MD, molecular dynamics; ROC, receiver operating characteristic; AUC, area under the ROC curve; ROC-EF, ROC enrichment factor; DUD-E, directory of useful decoys enhanced; MK2, MAP-kinase-activated protein kinase 2; CDK2, cyclin-dependent protein kinase 2; ROCK1, rho-associated protein kinase 1; AKT1, serine/threonine-protein kinase AKT1; IGF1R, insulin-like growth factor 1-receptor; ABL, nonreceptor tyrosine protein kinase ABL1

■ REFERENCES

- (1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discovery* **2010**, *9*, 203–214.
- (2) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The Price of Innovation: New Estimates of Drug Development Costs. *J. Health Economics* **2003**, *22*, 151–185.
- (3) Lionta, E.; Spyrou, G.; Vassiliadis, D. K.; Cournia, Z. Structure-based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923–1938.
- (4) Lavecchia, A.; Di Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20*, 2839–2860.
- (5) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 98–104.
- (6) Ma, B.; Kumar, S.; Tsai, C. J.; Nussinov, R. Folding Funnel and Binding Mechanisms. *Protein Eng., Des. Sel.* **1999**, *12*, 713–720.
- (7) Carlson, H. A.; McCammon, J. A. Accommodating Protein Flexibility in Computational Drug Design. *Mol. Pharmacol.* **2000**, *57*, 213–218.
- (8) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598–1603.
- (9) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnel, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Protein: Struct., Funct., Genet.* **1995**, *21*, 167–195.
- (10) Harrison, S. C.; Durrant, R. Is There a Single Pathway for the Folding of a Polypeptide Chain? *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 4028–4030.
- (11) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.
- (12) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 186–193.
- (13) Korb, O.; Olsson, T. S.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and Limitations of Ensemble Docking. *J. Chem. Inf. Model.* **2012**, *52*, 1262–1274.
- (14) Huang, S. Y.; Zou, X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Protein: Struct., Funct., Genet.* **2007**, *66*, 399–421.
- (15) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving Database Enrichment through Ensemble Docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621–627.
- (16) Barril, X.; Morley, S. D. Unveiling the Full Potential of Flexible Receptor Docking Using Multiple Crystallographic Structures. *J. Med. Chem.* **2005**, *48*, 4432–4443.
- (17) Park, S. J.; Kufareva, I.; Abagyan, R. Improved Docking, Screening and Selectivity Prediction for Small Molecule Nuclear Receptor Modulators Using Conformational Ensembles. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 459–471.
- (18) Damm, K. L.; Carlson, H. A. Exploring Experimental Sources of Multiple Protein Conformations in Structure-based Drug Design. *J. Am. Chem. Soc.* **2007**, *129*, 8225–8235.
- (19) Isvoran, A.; Badel, A.; Craescu, C. T.; Miron, S.; Miteva, M. A. Exploring NMR Ensembles of Calcium Binding Proteins: Perspectives to Design Inhibitors of Protein-Protein Interactions. *BMC Struct. Biol.* **2011**, *11*, 24–35.
- (20) Osguthorpe, D. J.; Sherman, W.; Hagler, A. T. Generation of Receptor Structural Ensembles for Virtual Screening Using Binding Site Shape Analysis and Clustering. *Chem. Biol. Drug Des.* **2012**, *80*, 182–193.
- (21) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based Virtual Screening Reveals Potential Novel Antiviral Compounds for Avian Influenza Neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878–3894.
- (22) Amaro, R. E.; Schnauffer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A. Discovery of Drug-like Inhibitors of an Essential RNA-editing Ligase in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17278–17283.
- (23) Durrant, J. D.; Hall, L.; Swift, R. V.; Landon, M.; Schnauffer, A.; Amaro, R. E. Novel Naphthalene-based Inhibitors of *Trypanosoma brucei* RNA Editing Ligase 1. *PLoS Neglected Trop. Dis.* **2010**, *4*, e803.
- (24) Wassman, C. D.; Baronio, R.; Demir, O.; Wallentine, B. D.; Chen, C. K.; Hall, L. V.; Salehi, F.; Lin, D. W.; Chung, B. P.; Hatfield, G. W.; Richard Chamberlin, A.; Luecke, H.; Lathrop, R. H.; Kaiser, P.; Amaro, R. E. Computational Identification of a Transiently Open L1/S3 Pocket for Reactivation of Mutant p53. *Nat. Commun.* **2013**, *4*, 1407–1415.

- (25) Demir, O.; Labaied, M.; Merritt, C.; Stuart, K.; Amaro, R. E. Computer-aided Discovery of Trypanosoma brucei RNA-editing Terminal Uridyl Transferase 2 Inhibitors. *Chem. Biol. Drug Des.* **2014**, *84*, 131–139.
- (26) Kiss, R.; Jojart, B.; Schmidt, E.; Kiss, B.; Keseru, G. M. Identification of Novel Histamine H4 Ligands by Virtual Screening on Molecular Dynamics Ensembles. *Mol. Inf.* **2014**, *33*, 264–268.
- (27) Ivetac, A.; Swift, S. E.; Boyer, P. L.; Diaz, A.; Naughton, J.; Young, J. A. T.; Hughes, S. H.; McCammon, J. A. Discovery of Novel Inhibitors of HIV-1 Reverse Transcriptase Through Virtual Screening of Experimental and Theoretical Ensembles. *Chem. Biol. Drug Des.* **2014**, *83*, 521–531.
- (28) Wang, B.; Buchman, C. D.; Li, L.; Hurley, T. D.; Meroueh, S. O. Enrichment of Chemical Libraries Docked to Protein Conformational Ensembles and Application to Aldehyde Dehydrogenase 2. *J. Chem. Inf. Model.* **2014**, *54*, 2105–2116.
- (29) Tarcsay, A.; Paragi, G.; Vass, M.; Jojart, B.; Bogar, F.; Keseru, G. M. The Impact of Molecular Dynamics Sampling on the Performance of Virtual Screening Against GPCRs. *J. Chem. Inf. Model.* **2013**, *53*, 2990–2999.
- (30) Nichols, S. E.; Baron, R.; Ivetac, A.; McCammon, J. A. Predictive Power of Molecular Dynamics Receptor Structures in Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 1439–1446.
- (31) Xu, M.; Lill, M. A. Utilizing Experimental Data for Reducing Ensemble Size in Flexible-protein Docking. *J. Chem. Inf. Model.* **2012**, *52*, 187–198.
- (32) Johnson, L. N.; Lewis, R. J. Structural Basis for Control by Phosphorylation. *Chem. Rev.* **2001**, *101*, 2209–2242.
- (33) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912–1934.
- (34) Adams, J. A. Kinetic and Catalytic Mechanisms of Protein Kinases. *Chem. Rev.* **2001**, *101*, 2271–2290.
- (35) Ma, W. W.; Adjei, A. A. Novel Agents on the Horizon for Cancer Therapy. *Ca-Cancer J. Clin.* **2009**, *59*, 111–137.
- (36) Huang, M.; Shen, A.; Ding, J.; Geng, M. Molecularly Targeted Cancer Therapy: Some Lessons from the Past Decade. *Trends Pharmacol. Sci.* **2014**, *35*, 41–50.
- (37) Sun, C.; Bernards, R. Feedback and Redundancy in Receptor Tyrosine Kinase Signaling: Relevance to Cancer Therapies. *Trends Biochem. Sci.* **2014**, *39*, 465–474.
- (38) Kikuchi, R.; Nakamura, K.; MacLauchlan, S.; Ngo, D. T.; Shimizu, I.; Fuster, J. J.; Katanasaka, Y.; Yoshida, S.; Qiu, Y.; Yamaguchi, K.; Matsushita, T.; Murohara, T.; Gokce, N.; Bates, D. O.; Hamburg, N. M.; Walsh, K. An Antiangiogenic Isoform of VEGF-A Contributes to Impaired Vascularization in Peripheral Artery Disease. *Nat. Med.* **2014**, *20*, 1464–1471.
- (39) Muth, F.; Gunther, M.; Bauer, S. M.; Doring, E.; Fischer, S.; Maier, J.; Druckes, P.; Koppler, J.; Trappe, J.; Rothbauer, U.; Koch, P.; Lauffer, S. A. Tetra-substituted Pyridinylimidazoles as Dual Inhibitors of p38alpha Mitogen-activated Protein Kinase and c-Jun N-terminal Kinase 3 for Potential Treatment of Neurodegenerative Diseases. *J. Med. Chem.* **2015**, *58*, 443–456.
- (40) Clark, J. D.; Flanagan, M. E.; Telliez, J. Discovery and Development of Janus Kinase (JAK) Inhibitors for Inflammatory Diseases. *J. Med. Chem.* **2014**, *57*, 5023–5038.
- (41) Barnes, P. J. New Anti-inflammatory Targets for Chronic Obstructive Pulmonary Disease. *Nat. Rev. Drug Discovery* **2013**, *12*, 543–559.
- (42) Banks, A. S.; et al. An ERK/Cdk5 Axis Controls the Diabetogenic Actions of PPARγ. *Nature* **2014**, *517*, 391–395.
- (43) Kinch, M. S. An Analysis of FDA-approved Drugs for Oncology. *Drug Discovery Today* **2014**, *19*, 1831–1835.
- (44) Wu, P.; Nielsen, T. E.; Clausen, M. H. FDA-approved Small-molecule Kinase Inhibitors. *Trends Pharmacol. Sci.* **2015**, *36*, 422–439.
- (45) Lamontanara, A. J.; Gencer, E. B.; Kuzyk, O.; Hantschel, O. Mechanisms of Resistance to BCR-ABL and other Kinase Inhibitors. *Biochim. Biophys. Acta, Proteins Proteomics* **2013**, *1834*, 1449–1459.
- (46) Ma, L.; Shan, Y.; Bai, R.; Xue, L.; Eide, C. A.; Ou, J.; Zhu, L. J.; Hutchinson, L.; Cerny, J.; Khoury, H. J.; Sheng, Z.; Druker, B. J.; Li, S.; Green, M. R. A Therapeutically Targetable Mechanism of BCR-ABL-Independent Imatinib Resistance in Chronic Myeloid Leukemia. *Sci. Transl. Med.* **2014**, *6*, 252ra121.
- (47) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.
- (48) Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Eyck, L. F. Surface Comparison of Active and Inactive Protein Kinases Identifies a Conserved Activation Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 17783–17788.
- (49) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (50) Schrödinger Release 2014-4: *Schrödinger Suite 2014-4 Protein Preparation Wizard; Epik version 3.0*, S.; LLC, New York, NY, 2014. *Impact version 6.5*; Schrödinger, LLC, New York, NY, 2014. *Prime version 3.8*; Schrödinger, LLC, New York, NY, 2014.
- (51) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A Hierarchical Approach to All-atom Protein Loop Prediction. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 351–367.
- (52) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. *J. Mol. Biol.* **2002**, *320*, 597–608.
- (53) Schrödinger Release 2014-4: *Prime, v.*; Schrödinger, LLC, New York, NY, 2014.
- (54) Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein pKa Values. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 704–721.
- (55) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein-ligand Complexes. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 765–783.
- (56) Olsson, M. H.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.
- (57) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.
- (58) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198–210.
- (59) Case, D. A.; Betz, R. M.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Monard, G.; Needham, P.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Salomon-Ferrer, R.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; York, D. M.; Kollman, P. A. *AMBER 2015*; University of California, San Francisco, 2015.
- (60) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (61) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.
- (62) Joung, I. S.; Cheatham, T. E., 3rd Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.
- (63) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (64) Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations

- with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.
- (65) Salomon-Ferrer, R.; Gotz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.
- (66) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (67) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem., Int. Ed.* **1999**, *38*, 236–240.
- (68) Metwally, E.; Ismail, H. A.; Davison, J. S.; Mathison, R. A Tree-based Algorithm for Determining the Effects of Solvation on the Structure of Salivary Gland Tripeptide NH3+-D-PHE-D-GLU-GLY-COO. *Biophys. J.* **2003**, *85*, 1503–1511.
- (69) Durrant, J. D.; Votapka, L.; Sorensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory Comput.* **2014**, *10*, 5047–5056.
- (70) Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess, B.; Team, G. d. *GROMACS User Manual*, version 5.0.3; 2014.
- (71) Grant, B. J.; Rodrigues, A. P.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. Bio3d: An R Package for the Comparative Analysis of Protein Structures. *Bioinformatics* **2006**, *22*, 2695–2696.
- (72) Schrödinger Release 2014-4: *LigPrep*, v.; Schrödinger, LLC, New York, NY, 2014.
- (73) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (74) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.
- (75) Small-Molecule Drug Discovery Suite 2015-4: *Glide*, v.; Schrödinger, LLC, New York, NY, 2015.
- (76) Metz, C. E. Basic Principles of ROC Analysis. *Semin. Nucl. Med.* **1978**, *8*, 283–298.
- (77) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874.
- (78) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- (79) Nicholls, A. What do we Know and When do we Know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
- (80) Swift, R. V.; Jusoh, S. A.; Offutt, T. L.; Li, E. S.; Amaro, R. E. Knowledge-Based Methods to Train and Optimize Virtual Screening Ensembles. *J. Chem. Inf. Model.* **2016**, *56*, 830–842.
- (81) Amaro, R. E.; Li, W. W. Emerging Methods for Ensemble-based Virtual Screening. *Curr. Top. Med. Chem.* **2010**, *10*, 3–13.
- (82) Ellingson, S. R.; Miao, Y.; Baudry, J.; Smith, J. C. Multi-conformer Ensemble Docking to Difficult Protein Targets. *J. Phys. Chem. B* **2015**, *119*, 1026–1034.
- (83) Damm, K. L.; Carlson, H. A. Exploring Experimental Sources of Multiple Protein Conformations in Structure-Based Drug Design. *J. Am. Chem. Soc.* **2007**, *129*, 8225–8235.
- (84) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of Molecular Docking to Induced Fit Effects in Influenza Virus Neuraminidase. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 855–869.
- (85) Yoon, S.; Welsh, W. J. Identification of a Minimal Subset of Receptor Conformations for Improved Multiple Conformation Docking and Two-step Scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.
- (86) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-ligand Docking Against Non-native Protein Conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214–2225.

Supporting Information

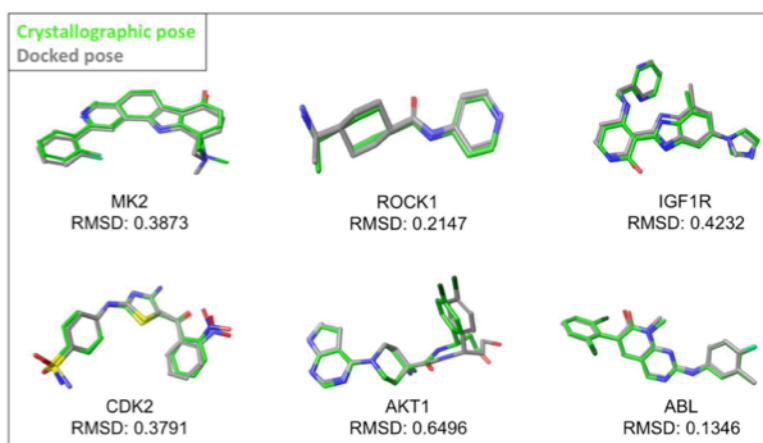
**Enhancing Virtual Screening Performance of Protein Kinases with
Molecular Dynamics Simulations**

Tavina L. Offutt, Robert V. Swift, Rommie E. Amaro*

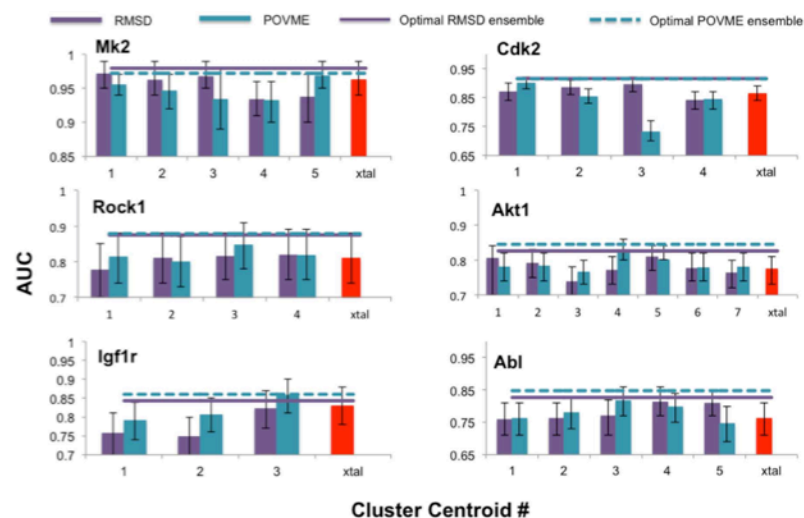
Department of Chemistry and Biochemistry, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92092-0340, United States

*Corresponding author: ramaro@ucsd.edu

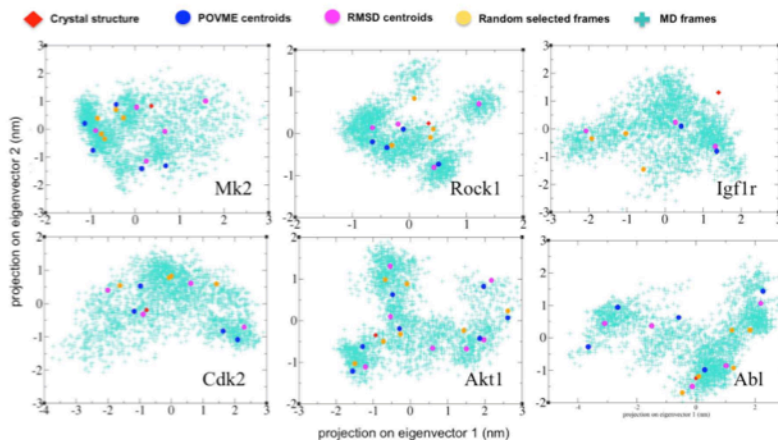
SUPPORTING FIGURES



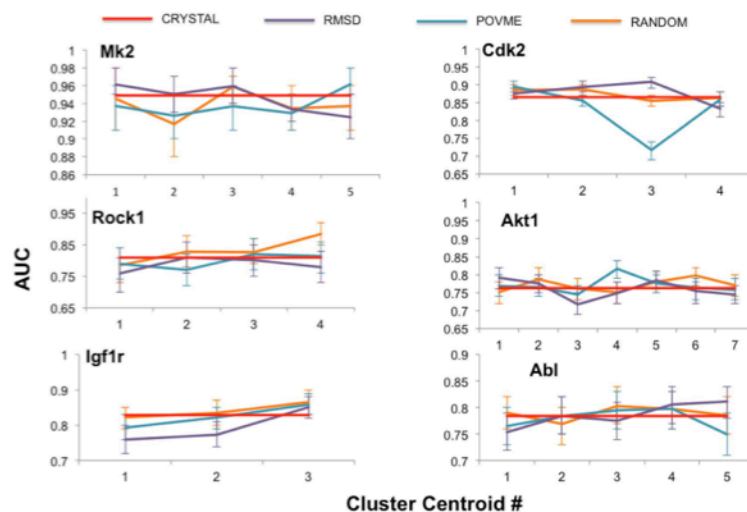
Supporting Figure S1. Comparison between the crystallographic and docked pose of the co-crystallized inhibitors. The inhibitors are colored by atom type (C: green for crystallographic pose, and gray for the docked pose; N: blue; O: red; S: yellow; F: light green; Cl: dark green). The RMSD between each heavy atom pair was calculated using Schrödinger's Maestro superposition tool.



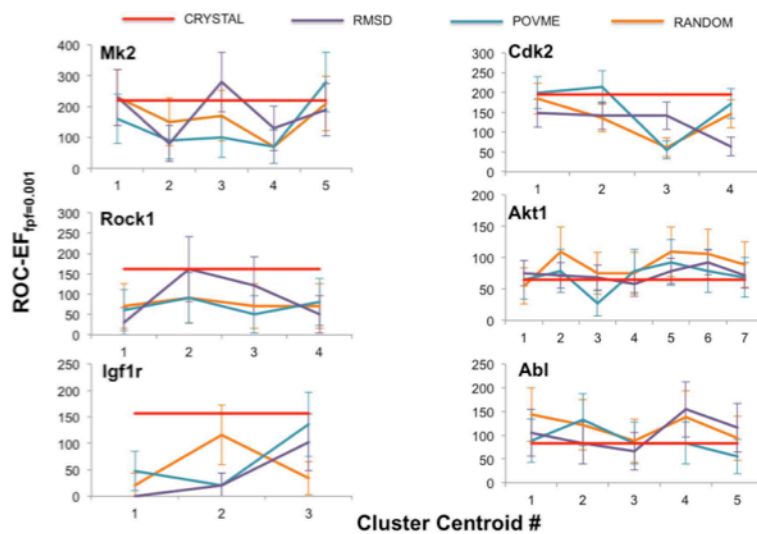
Supporting Figure S2. The cluster centroids and crystal's AUC against the training set for each protein kinase. The AUC of the optimal trained ensembles (the ensemble with the highest AUC) using RMSD and POVME centroids are shown for comparison.



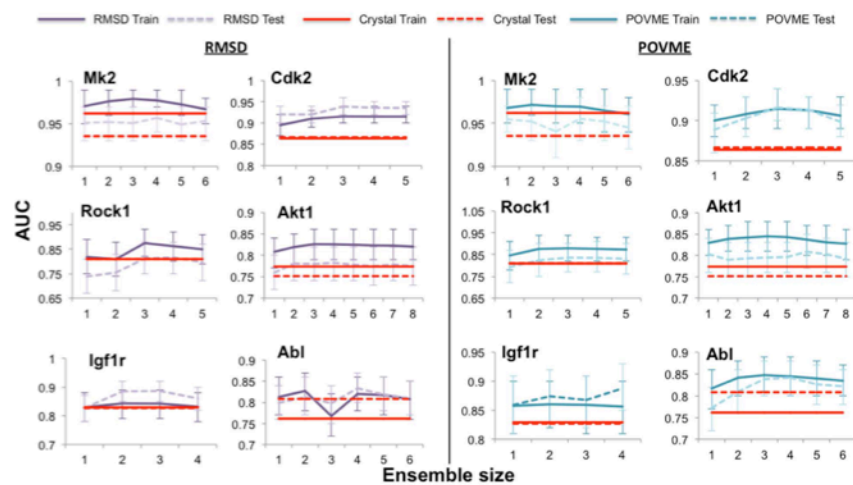
Supporting Figure S3. Comparison between all structural selection methods. Principal components 1 and 2 of the MD snapshots, cluster representatives (centroids), randomly selected MD frames, and the crystal structure projected onto 2D space for each protein kinase are shown.



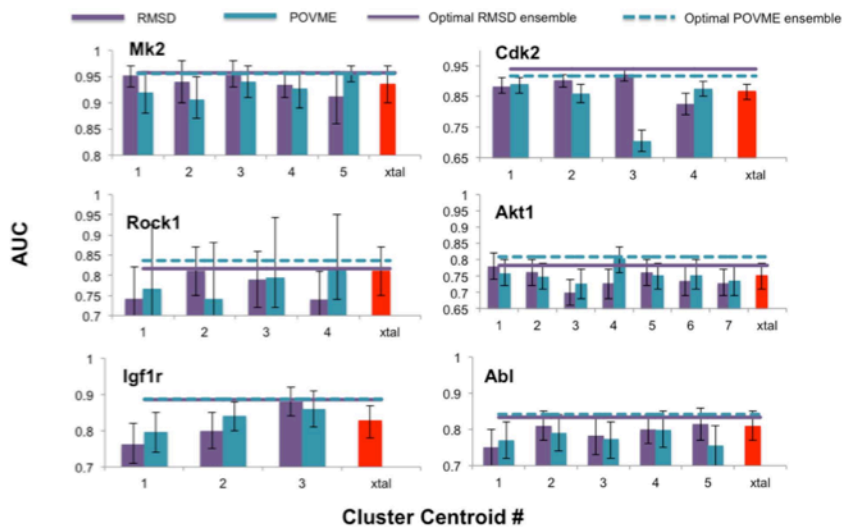
Supporting Figure S4. The RMSD and POVME cluster centroids and randomly selected frames AUC values against the entire dataset of actives and decoys. The AUC of the crystal structure is shown for comparison.



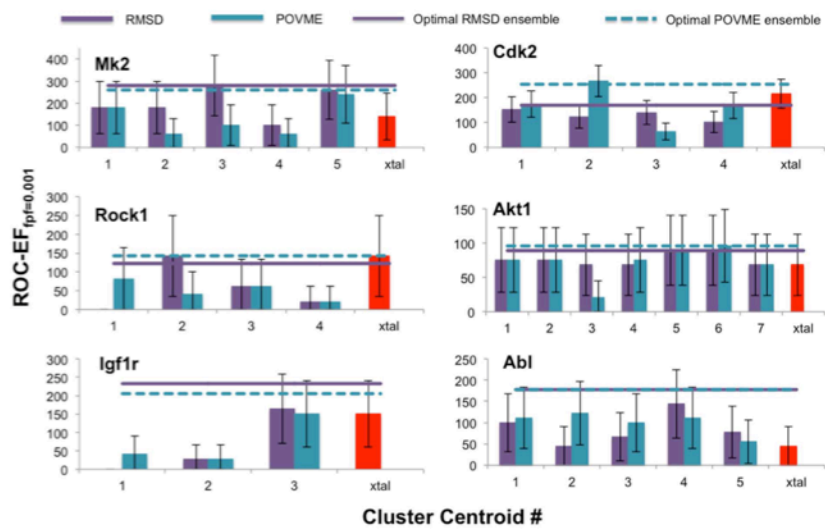
Supporting Figure S5. The RMSD and POVME cluster centroids and randomly selected frames ROC-EF values against the entire dataset of actives and decoys. The ROC-EF of the crystal structure is shown for comparison.



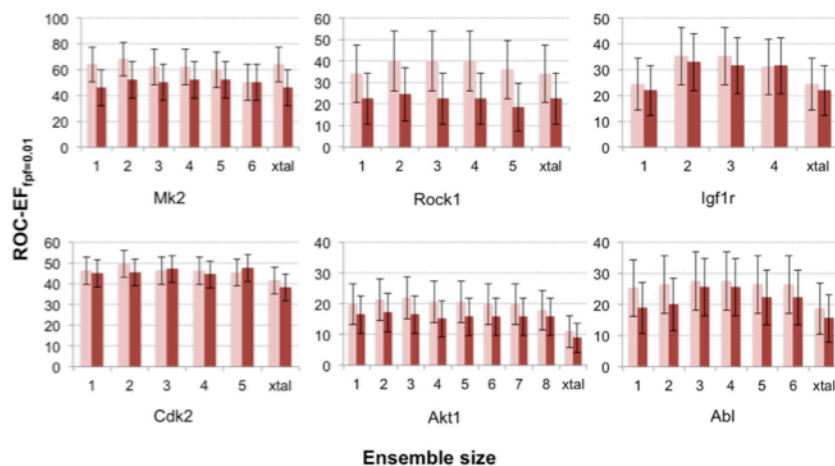
Supporting Figure S6. The trained ensemble sizes and crystal structures AUC values against the training and test set across all six protein kinases are shown. The 95% confidence intervals overlap between the training and test set, validating the training method.



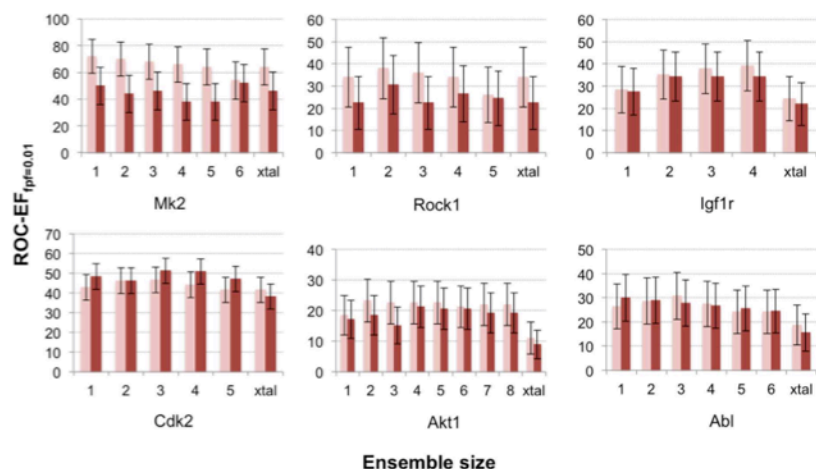
Supporting Figure S7. The cluster centroids and crystal's AUC against the test set for each protein kinase. The AUC of the optimal trained ensembles (the ensemble with the highest AUC) using RMSD and POVME centroids are shown for comparison.



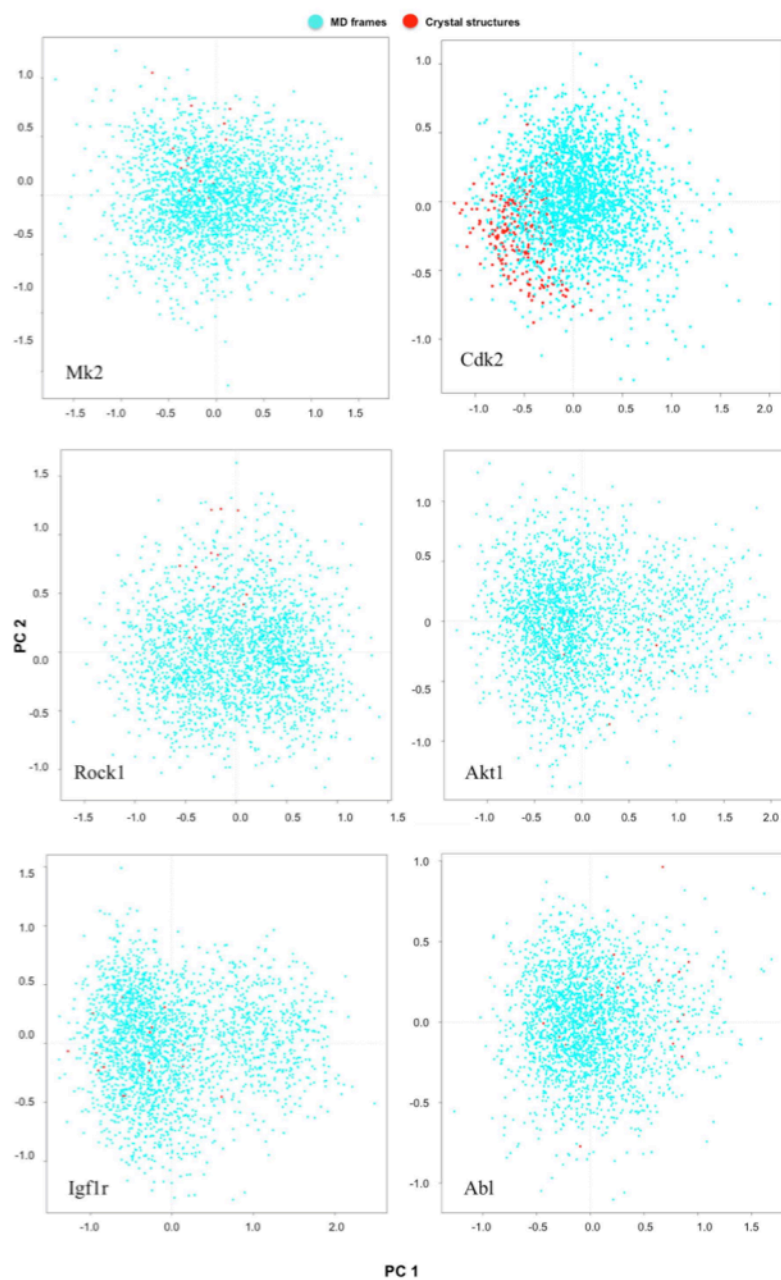
Supporting Figure S8. The cluster centroids and crystal's ROC-EF against the test set for each protein kinase. The ROC-EF of the optimal trained ensemble (the ensemble with the highest ROC-EF) using RMSD and POVME centroids are shown for comparison.



Supporting Figure S9. ROC-EF values at a later false positive fraction (fpf=0.01) against the training and test set for each ensemble combination using RMSD centroids across all six protein kinases. The training set is shown as light red bars and the test set is shown as dark red bars. The 95% confidence intervals are shown above each bar.



Supporting Figure S10. ROC-EF values at a later false positive fraction (fpf=0.01) against the training and test set for each ensemble combination using POVME centroids across all six protein kinases. The training set is shown as light red bars and the test set is shown as dark red bars. The 95% confidence intervals are shown above each bar.



Supporting Figure S11. Comparison between Crystal Structures and MD Conformations. Principal components 1 and 2 of the binding site for all available crystal structures and the MD frames are shown.

Supporting Table S1. Number of Crystal Structures used for PCA Comparison between MD and Crystal Structure Conformations

Protein Kinase	Number of Available Crystal Structures ^a	Number of Structures Used after Filtering Criteria ^b
MK2	18	16
CDK2	369	234
ROCK1	18	13
AKT1	31	11
IGF1R	21	20
ABL	49	20

^aA PDB search using the UniProt ID and a BLAST search in the Bio3D R package was used to find all crystal structures for all six protein kinases.

^bOnly inhibitor-bound crystal structures were extracted from the PDB and BLAST search results. Within the inhibitor-bound structures, inhibitor that were bound to an alternate site were also eliminated.

Chapter 3, in full, is a reprint of the material as it appears in Enhancing Virtual Screening Performance of Protein Kinases with Molecular Dynamics Simulations 2016. Offutt, Tavina L.; Swift, Robert, V.; Amaro, Rommie E., J Chem Inf Mod, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Knowledge-Based Methods to Train and Optimize Virtual Screening Ensembles



Knowledge-Based Methods To Train and Optimize Virtual Screening Ensembles

Robert V. Swift,[†] Siti A. Jusoh,[‡] Tavina L. Offutt,[†] Eric S. Li,[†] and Rommie E. Amaro^{*,†}

[†]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093-0340, United States

[‡]Faculty of Pharmacy, Universiti Teknologi MARA, 42300 Bandar Puncak Alam, Malaysia

Supporting Information

ABSTRACT: Ensemble docking can be a successful virtual screening technique that addresses the innate conformational heterogeneity of macromolecular drug targets. Yet, lacking a method to identify a subset of conformational states that effectively segregates active and inactive small molecules, ensemble docking may result in the recommendation of a large number of false positives. Here, three knowledge-based methods that construct structural ensembles for virtual screening are presented. Each method selects ensembles by optimizing an objective function calculated using the receiver operating characteristic (ROC) curve: either the area under the ROC curve (AUC) or a ROC enrichment factor (EF). As the number of receptor conformations, N , becomes large, the methods differ in their asymptotic scaling. Given a set of small molecules with known activities and a collection of target conformations, the most resource intensive method is guaranteed to find the optimal ensemble but scales as $O(2^N)$. A recursive approximation to the optimal solution scales as $O(N^2)$, and a more severe approximation leads to a faster method that scales linearly, $O(N)$. The techniques are generally applicable to any system, and we demonstrate their effectiveness on the androgen nuclear hormone receptor (AR), cyclin-dependent kinase 2 (CDK2), and the peroxisome proliferator-activated receptor δ (PPAR- δ) drug targets. Conformations that consisted of a crystal structure and molecular dynamics simulation cluster centroids were used to form AR and CDK2 ensembles. Multiple available crystal structures were used to form PPAR- δ ensembles. For each target, we show that the three methods perform similarly to one another on both the training and test sets.



INTRODUCTION

Virtual screening (VS) is a valuable hit discovery tool with tremendous potential to improve the efficiency and reduce the costs of modern high throughput screens (HTS). Despite the increasing trend toward miniaturization and greater well plate density, reagents and other consumables drive up HTS costs, particularly when large corporate or commercial databases are screened.¹ Rationally prioritizing compounds for experimental testing can reduce costs. For example, during a virtual high throughput screen, a computational model is developed and applied to rank compounds for testing.^{2,3} When paired with high quality compound libraries, carefully constructed computational models can generate hit rates many fold above random.^{4,5} This can result in novel, structurally diverse actives from which several lead series can be selected. Structural diversity ultimately helps circumvent ADME-Tox and patent liabilities that can increase lead optimization costs.^{6,7}

Computational virtual screening models primarily fall into two classes, ligand-based⁸ and structure based, or docking methods.⁹ Ligand-based methods predict the activity of novel compounds by assessing their similarity to known actives. Docking methods, on the other hand, use predicted interactions between a small molecule and a target receptor to predict activity. Numerous benchmarking studies have reported that ligand-based methods yield greater hit rates

than structure-based methods.^{5,10,11} However, a reliance on chemical similarity may limit their ability to identify novel chemical matter. In contrast, the diversity of actives determined using docking methods is only constrained by the shape of the receptor-binding pocket. In principal, docking can enable the discovery of actives more novel and diverse than ligand-based methods. Consistently, numerous successful examples of docking in early stage discovery can be found in the literature.^{12–14}

Despite these successes, docking has traditionally been to a single static representation of the target. This static view is far from reality. In solution, a drug target is highly dynamic, and two notable models have been advanced that suggest a tight coupling between protein motion and small molecule binding. In 1958, Koshland proposed the induced fit model, which suggests that ligand binding induces a conformational change of the protein that enhances its affinity for the ligand.¹⁵ Conformational selection is a more recent explanation of small molecule binding that incorporates energy landscape theory.¹⁶ It proposes that binding stabilizes one of many preexisting conformers of the unbound target.^{17,18} Both models imply that the collection of low-energy receptor conformers

Received: November 12, 2015

Published: April 20, 2016

defining the bound state depend upon ligand identity; by extension, successful docking requires the receptor to be in, or at least near, the appropriate ligand-dependent bound state. Ensemble docking, in which each ligand is docked to a set of receptor conformers, was introduced in an effort to address this requirement.¹⁹

There are a variety of means to generate structures for ensemble docking, including crystallography²⁰ and NMR²¹ techniques. However, while experimental methods have shown promise, the materials, time, and expertise required to determine multiple, high quality structures is a significant bottleneck. In contrast, molecular dynamics (MD) simulations offer a relatively inexpensive alternative to generate diverse, realistic conformational states. This is largely the result of the recent implementation of MD codes on commodity graphical processor units (GPUs)^{22,23} and the dramatic speedup of simulation benchmarks.

Regardless of whether structures are generated by experiment or simulation, for ensemble docking to be successful, a subset of conformations likely to offer the best VS performance must be identified. Though several studies have provided hints,^{24–26} others have been unable to determine a meaningful relationship between observable receptor characteristics and virtual screening performance.^{27,28} Even with insights from a growing body of careful studies, it remains difficult or impossible to know *a priori* which receptor conformations will result in an ensemble with virtual screening utility.

The difficulties of selecting effective virtual screening conformations are compounded by the combinatorial nature of the ensemble selection process. When the number of receptor conformations is large, the problem results in a significant number of possibilities, and it can be difficult or impossible to know which of these ensembles will produce the best virtual screening performance.

Though systematic training and data-fusion methods exist that address similar issues in ligand-based VS, there is a relative paucity of knowledge-based structural selection methods. Despite this, other knowledge-based ensemble selection methods have been described in the literature. For example, Yoon and Welsh²⁹ proposed an ensemble docking method in which ensemble members are selected to maximize the correlation between the experimental and predicted binding affinities. The combinatorial problem was addressed by assigning each compound an ensemble score that consisted of a linear combination of score weights to each receptor conformation using a Monte Carlo scheme. Using estrogen receptor α , they demonstrated that the approach leads to more accurate classification than docking to the crystal structure alone.

While Yoon's and Welsh's method can produce stronger correlation with experimental binding affinities and result in enhanced VS performance, experimental binding measurements are required. This precludes the use of single-point HTS data and limits the method to compounds whose binding affinities have been measured or to those with dose–response curves, from which binding affinities may be inferred.

Rather than optimizing the correlation with experimental binding affinities, selecting ensembles to maximize the value of a binary classification metric offers greater flexibility. Since binary classification is categorical, once an appropriate activity threshold has been determined, any assay that delivers an activity measurement can be used. This opens the door to the use of single-point data, which is less expensive to determine

and typically can be found in greater abundance than careful binding affinity measurements.

For example, following a slightly different approach, Xu and Lill developed a knowledge-based ensemble selection technique that can be used with any type of affinity measurement.³⁰ In it, receptor conformers are first ranked by their ability to separate the average docking scores of active and inactive compounds. Then, by assuming that effective ensembles must be constructed from effective conformations, ensembles of successively larger size are formed by aggregating conformers from highest to lowest rank. While the assumption avoids the combinatorial problem, its severity went unexamined. For example, does the procedure ignore ensembles with significantly greater classification power? While the underlying assumption went unexamined, the approach appeared promising. When classification ability was examined as a function of ensemble size, the performances of the trained ensembles were comparable or better than the those of ensembles selected by aggregating structurally diverse receptor conformers.

A final approach, developed and widely applied by the Cavasotto and Abagyan groups, utilizes virtual screening performance on a small training set to select the most promising structure from an ensemble generated using either Monte Carlo side-chain sampling or normal-mode analysis.³¹ By including a ligand with the desired properties, for example, a high affinity binder or a receptor agonist/antagonist, the search may be biased toward structures that enrich ligands with similar properties. During model generation, the VS ability of each target conformer is evaluated, and conformational sampling continues until VS performance converges. Following convergence, a single best performing structure can be derived and used for cross docking, selectivity studies, or VS. Alternatively, multiple conformers may be extracted and combined into useful ensembles, and the methods we introduce here may prove useful in such an approach.

In this work, we present three new training methods that select structure-based ensembles for VS use. All three methods construct ensembles by optimizing one of two binary classification metrics, which makes them flexible and enables their use with single-point data, competition assay data (e.g., IC₅₀ values), or other binding data. To address the combinatorial problem, the population of ensembles is generated by complete enumeration, and two different heuristics are designed to generate population samples biased to exclude low performing ensembles. These approaches lead to different asymptotic scaling as the number of receptor conformations becomes large, and they allow us to examine the severity of the approximations underlying each heuristic relative to the enumerative solution.

Each method is evaluated on three different target proteins with active and decoy molecules taken from the DUD-E:³² the androgen nuclear hormone receptor (AR), the cyclin-dependent kinase 2 (CDK2), and the peroxisome proliferator-activated receptor δ (PPAR- δ). Target conformations were selected from a range of sources, including RMSD and volumetric clustering of conventional MD simulations, as well as multiple crystal structures.

■ METHODS

Data Sets and Target Proteins. The knowledge-based training methods were tested on three protein targets: the androgen receptor, the cyclin-dependent kinase 2 (CDK2), and the peroxisome proliferator-activated receptor δ (PPAR- δ).

Conformations generated by volumetric clustering of conventional MD trajectories along with the crystal structure PDBID 2AM9 were used to train androgen receptor ensembles. Similarly, conformations generated by RMSD clustering of MD trajectories, along with the crystal structure PDBID 4GCJ, were used to train ensembles of CDK2. Clustering and simulation details are provided in subsequent sections. For PPAR- δ , ensemble training was performed using 12 crystal structure conformations with the following PDBIDs (Uniprot Q03181): 2AWH, 2B50, 2J14, 2Q5G, 2XYJ, 2ZNP, 3DY6, 3ET2, 3GZ9, 3PEQ, 3SP9, and 3TKM. Structures were selected to ensure a resolution of 3.0 Å or lower and to ensure that each ligand was unique. Additionally, all of the structures are antagonist bound, which is consistent with the antagonists that make up the actives of the training and test sets, as described below.

Active and decoy ligand sets from the Directory of Useful Decoys-Ehanced (DUD-E)³² were used to perform virtual screening for each target. While a complete description of ligand set curation can be found in the original reference, we briefly describe the process here. Compounds in ChEMBL whose affinities (IC_{50} , EC_{50} , K_i , K_d) were less than or equal to 1 μ M were clustered by their Bemis Murcko (BM) frameworks.³³ Compounds with the highest affinity from each cluster were pooled and resulted in sets of actives with unique BM frameworks. For each active, 50 decoys were selected from the Zinc database by matching the molecular weight, logP, number of rotatable bonds, hydrogen bond donor and acceptor counts, and net formal charge (determined in a pH range from 6 to 8) of the active. To reduce the number of false negatives, only the 25% most dissimilar decoys, as judged by Tanimoto scores using ECF4P fingerprints, were retained.

Evaluating the classification performance of a knowledge-based model on the training set will generally provide an overly optimistic estimate of the model's ability to correctly distinguish active and inactive molecules.³⁴ To provide a more realistic estimate of the trained model's classification ability, DUD-E compounds were randomly split in half, while maintaining the decoy-to-active ratio, forming training and test sets. The androgen receptor training and test sets were composed of 7150 compounds, 133 of which were active compounds. The CDK2 training and test sets were composed of 14,162 compounds, 237 of which were active compounds. The PPAR- δ training and test sets were composed of 6245 compounds, 120 of which were active.

Molecular Dynamics. Except as noted, CDK2 and androgen simulations were performed identically. Simulations were initiated from a crystal structure of either the androgen receptor (PDBID 2AM9) or CDK2 (PDBID 4GCJ). The sulfate ion, glycerol, and the dithiothreitol molecule were deleted from 2AM9, while four molecules of ethanediol were deleted from 4GCJ. In 2AM9, K836, K846, N848, and E893 are far from the receptor-binding pocket and have unresolved side-chain atoms. Schrödinger's Prime^{35,36} was used to add them. In 4GCJ, atoms from the following residues had multiple occupancy values: D38, S46, D127, K129, R169, L212, M233, K237, K250, S264, and H268. In each case, the position with the larger value was retained. Protonation states for both 2AM9 and 4GCJ were predicted at pH 7.0 using the program PROPKA,^{37–39} and hydrogen atom positions were assigned and optimized using Schrödinger's Protein Preparation Wizard. Following protonation, water molecules with fewer than three hydrogen bonds to nonwater molecules were removed. The

protonated crystal structures were built for MD simulation using the xLEaP program that accompanies AMBER14.⁴⁰ The cholesterol and RC-3-89 ligand parameters were generated using the Antechamber program in AMBER14. Ligand atomic partial charges were determined from the crystallographic conformations using the AM1-BCC method,⁴¹ and all other force field terms were assigned according to the generalized Amber force field (GAFF).⁴² Each system was immersed in a box of pre-equilibrated TIP4PEW water⁴³ that provided a minimum 10 Å water pad between the protein and the boundary of the periodic box in the x -, y -, and z -directions. Each system was brought to electric neutrality by the addition of an appropriate number of chloride or sodium ions, modeled using the parameters developed by Joung and Cheatham.⁴⁴ The androgen receptor system was comprised of 54,014 atoms, and the CDK2 system was composed of 50,644 atoms. The potential energy was described by the AMBER14 force field with the Stony Brook correction.⁴⁵ A 20,000-step minimization was performed with 2 kcal mol⁻¹ Å⁻² heavy atom backbone restraint in two stages. During the first step, a 19,500-steepest descent minimization was conducted. The second step entailed a 500-step conjugate gradient minimization. Following minimization, a 200 ps NPT simulation was carried out at 300 K and 1 atm. Pressure was maintained with a Monte Carlo barostat with 100 steps between volume changes and a pressure relaxation time of 2 ps⁻¹. Following the NPT simulation, a 5 ns NVT simulation was conducted, and restart files were written every 1 ns. These restart files were used to initiate five 20 ns NVT simulations, and frames were written every 2000 fs. All 50,000 frames were concatenated yielding a 100 ns trajectory. During NPT and all NVT simulations, hydrogen heavy atom bonds were constrained using the SHAKE algorithm,⁴⁶ and a 2 fs time step was used. Temperature was maintained in all simulations using a Langevin thermostat with a collision frequency of 2 ps⁻¹. The Particle Mesh Ewald method was used to treat long-range electrostatics,²³ and simulations were performed using pmemd.cuda on a GeForce GTX TITAN card from NVIDIA. During NVT production runs, the simulation setup resulted in an average timing of 30.29 ns/day on the androgen receptor system and 31 ns/day on the CDK2 system.

Binding Site Clustering. The 100 ns trajectories were each subsampled at an interval of 40 ps, or every 20th frame, resulting in a total of 2500 frames, which were then clustered. Prior to clustering, external translation and rotation were removed from each trajectory by minimizing the RMSD distance of the C α backbone atoms to the equivalent atoms of the first sampled frame of the trajectory.

For the androgen receptor trajectory, the binding site shape of each sampled structure was determined using POVME 2.0.⁴⁷ Inclusion regions were autodetected using the testosterone ligand as input. Following binding site characterization, the Tanimoto volume overlap between all pairs of structures was calculated, from which a normalized volume overlap matrix was generated. Finally, hierarchical clustering was applied to the overlap matrix, 10 clusters were generated, and the structures corresponding to each of the cluster centroids were retained for docking. Ligand-based autodetection of inclusion regions and hierarchical clustering are features that will be released in the forthcoming version of POVME.

For the CDK2 trajectory, RMSD clustering was performed using the algorithm described in Daura et al.⁴⁸ as implemented in version 5.0.3 of the GROMACS `g_cluster` program.

Clustering was performed on the heavy atoms of all residues within 10 Å of the bound inhibitor RC-3-89 in the crystal structure PDBID 4GCJ. A cutoff of 1.6 Å resulted in five clusters, and cluster centroids were retained for docking.

Docking. The Glide SP algorithm, from Schrödinger, was used to perform docking to all target conformations.⁴⁹ The algorithm generates a series of ligand poses. Relative to the protein receptor, each pose has a unique position and orientation. Each pose is also distinguished by a unique conformation. Following generation, all poses are independently subjected to a set of hierarchical filters that utilize precomputed grids to estimate ligand–receptor interaction energies. In the initial filter, the steric complementarities of ligand poses with the receptor are computed using a grid-based version of ChemScore. Poses that pass the initial filter are minimized in a grid-based approximation of the OPLS pose–receptor interaction energy. Following minimization, Emodel, an empirical scoring function optimized to compare pose energetics, is used to identify the best pose for each ligand. Finally, a “docking score” is reported for each ligand. The docking score is an empirical ligand binding affinity estimate, which incorporates Epik state penalties that are based on the predicted populations of alternative ligand protonation and tautomerization states.⁵⁰

Prior to docking, two-dimensional representations of active and decoy molecules were downloaded from the DUD-E in SDF format. Schrödinger’s LigPrep program⁵¹ was used to add hydrogen atoms and generate three-dimensional ligand structures. Alternative protonation and tautomer states were determined at pH 7 using the Epik program, with default settings. Alternative ring conformations were not generated since these are produced by Glide during docking. Input chiralities were retained, and all other options were set to their default values.

Receptor conformations were prepared for docking as follows. TIP4PEW water and chloride ions were removed from the MD trajectory. The resulting trajectory, which consisted of either the androgen receptor and the testosterone ligand or CDK2 and the inhibitor RC-3-89 were clustered as described above, resulting in 10 and 6 cluster centroids, respectively. Schrödinger’s Protein Preparation Wizard was used to generate correct atom types for Glide grid generation. Atom coordinates were not altered in the process. Protonation states from the MD simulation were retained, and neither hydrogen bond network optimization nor structural minimizations were conducted. For each cluster centroid, the grid center was positioned on the center of geometry of the ligand; all other options were set to their default values.

Docking was performed using Glide with the SP scoring function. All other options were set to their default values. Docking was conducted locally on a Dell Precision T7500n workstation with a dual six-core Intel X5680 processor, and each compound required roughly 15 s to dock.

Performance Analysis. Receiver operating characteristic, or ROC, curves were used to evaluate the performance of each ensemble. ROC curves provide two useful measures of binary classification performance: the area under the curve (AUC) and the ROC enrichment factor (EF).

A ROC curve is determined by successively moving a threshold through compounds ranked by their docking scores. By assuming all compounds with scores better than the threshold are active, a true positive fraction (TPF) and false positive fraction (FPF) can be calculated at each threshold. For

example, the TPF is the fraction of active compounds whose docking scores are equal to or better than the threshold, ΔG_T . TPF can be calculated as an average over an indicator function, γ_i , as described by eq 1.

$$\text{TPF}(\Delta G_T) = \langle \gamma \rangle_A = \frac{1}{N_A} \sum_{i=1}^{N_A} \gamma_i$$

$$\text{where } \gamma_i = \begin{cases} 1 & \text{if } \Delta G_i \leq \Delta G_T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In eq 1, N_A is the total number of active compounds. For a given active, the indicator function γ takes a value of 1 if the value of the docking score, ΔG_i , is better than or identical to the threshold and a value of 0 otherwise. Similarly, the FPF is the fraction of inactive compounds whose docking scores are equal to or better than the threshold. It is also determined as an average of γ , but over the inactive compounds.

$$\text{FPF}(\Delta G_T) = \langle \gamma \rangle_I = \frac{1}{N_I} \sum_{i=1}^{N_I} \gamma_i$$

$$\text{where } \gamma_i = \begin{cases} 1 & \text{if } \Delta G_i \leq \Delta G_T \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

In eq 2, N_I is the total number of inactive compounds, and all other terms are defined identically to eq 1. Once the TPF and FPF values have been calculated at each threshold, they are plotted along the y-axis and x-axis, respectively, resulting in a ROC curve.

The area under the ROC curve (AUC) is equivalent to the probability that a virtual screening protocol will rank a randomly selected active compound ahead of a randomly selected inactive compound.⁵² An AUC value of 0.5 corresponds to random selection, or a method with no classification power, while a value of 1 indicates perfect separation of active and inactive compounds. Additionally, the value of the AUC is independent of the fraction of actives in the database, it has no free parameters, and an analytic estimate of its standard error is known.⁵² The AUC value can be estimated using a left-handed Riemann sum, which is equivalent to averaging the TPF values at each inactive compound of the ranked list.

$$\text{AUC} = \langle \text{TPF} \rangle_I \quad (3)$$

While the AUC is a useful measure of global classification performance, the early enrichment, or the preferential ranking of active compounds early in the ranked list, is often used to judge the quality of a virtual screen. Enrichment factors are frequently calculated as the ratio of the fraction of actives found in a given percent of the ranked database to the fraction of actives in the total database. Unfortunately, the maximum value of this popular metric depends on the ratio of inactive to active compounds in the screened database.⁵² This makes retrospective method comparison difficult. To circumvent this complicating factor, we use the so-called “ROC enrichment”, whose maximum value is independent of the ratio of decoy to active compounds. The ROC enrichment factor (EF) is the ratio of the TPF, determined at some FPF of interest, to the FPF of interest.⁵²

$$\text{EF}(\text{FPF}) = \frac{\text{TPF}(\text{FPF})}{\text{FPF}} \quad (4)$$

Random classification is indicated by an EF value of 1, and perfect separation of actives and decoys is given by a maximum value of PPF^{-1} . Like the AUC, the standard error of the ROC enrichment factor may be calculated analytically, which facilitates statistical analysis.⁵²

Statistical Analysis. For any VS protocol, classification performance will vary as a result of having different compounds in the screened database. Confidence intervals capture the magnitude of this variability. For example, assuming repeated screens are performed identically on different databases, the true mean should be found within identically constructed 95% confidence intervals (CI_{95}) in 95% of the measurements. CI_{95} values were constructed according to eq 5.⁵³

$$\text{CI}_{95} = l \times \text{SE} \quad (5)$$

The standard error, SE, of the calculated classification metric (AUC or EF) is given and is calculated differently for AUC and ROC-EF values. The exact form that each takes is provided in the Supporting Information. The value of l is selected such that $\pm l$ bounds 95% of Student's t -distribution, where the number of degrees of freedom was determined by subtracting one from the sum of the number of active and inactive compounds.

Ensemble Scoring. Several different methods for combining multiple docking scores into a single docking score have been suggested. Reported protocols include creating composite grids of all ensemble members,^{19,54} treating conformations as an independent variables during docking,^{55,56} and using different weighted averages, which include arithmetic¹⁷ and Boltzmann weighted averages,⁵⁷ as well as averages using weights determined by knowledge-based methods.²⁹ One simple approach, and the one used in this work, takes the best scoring function value across all ensemble members. For example, a compound docked to an ensemble composed of N protein conformations will have N docking scores, $\{\Delta G_1, \Delta G_2, \dots, \Delta G_N\}$, and the ensemble score of the compound is defined as the smallest score of the set, i.e., $\min\{\Delta G_1, \Delta G_2, \dots, \Delta G_N\}$. If a compound has more than one protonation or tautomer state, the state with the lowest docking score is retained.

RESULTS

Given an arbitrary collection of target conformations, it is difficult to know which set will result in the best VS performance. Here, we provide three knowledge-based methods designed to systematize the selection process: the exhaustive method, the slow heuristic method, and the fast heuristic method, which are each introduced below.

Knowledge-Based Ensemble Selection. In the "exhaustive" method, at each ensemble size, all combinatorial possibilities are enumerated, and the complete ensemble population is constructed. As shown in the Supporting Information, if N is the total number of target conformations considered, the enumerative approach generates $2^N - 1$ ensembles. Using big O notation, this is expressed as $O(2^N)$. For example, given three conformations, labeled A, B, and C, seven ensembles can be constructed: three of size one (A, B, C), three of size two (AB, AC, BC), and one of size three (ABC). Both AUC and EF values are used to rank the performance of each, and the best performing ensemble is retained. Thus, the exhaustive method generates the entire population of ensembles, performs a census, and only retains the individual member with the desired performance characteristics.

In the "slow heuristic" method, ensembles are assembled recursively. In the first step, the performance of each receptor conformation is considered individually, and the best performer becomes the first ensemble member. Next, the remaining receptor conformations are added in turn, forming a series of two-membered ensembles, and the best ensemble is retained. The process is repeated until all receptor conformations have been added to the ensemble, and the top performer of any size is identified. In the Supporting Information, we show that the slow heuristic method generates $N(N+1)/2$ ensembles. Using big O notation, this is expressed as $O(N^2)$. For example, given three conformations A, B, and C, three one-membered ensembles (A, B, C) will be considered, and the best performer will be retained. If B is the top performing one-member ensemble, then two two-membered ensembles (BA and BC) will be constructed, and one three-membered ensemble (ABC) will be constructed. Thus, the slow heuristic method is designed to construct a biased sample of the ensemble population that excludes individuals that do not contain the best performing ensembles of smaller sizes.

Like the slow heuristic method, the "fast heuristic" method also assembles ensembles recursively. First, the classification performance of each individual conformation is ranked by either AUC or EF. Ensembles of increasing size are then constructed by merging conformers of successively decreasing performance. The performance of each conformation is considered only once. To identify the ensemble that performs best, each of the resulting ensembles must also be evaluated once. Thus, for N conformers, $2N - 1$ performance evaluations are required, and the scaling is linear. Using big O notation, this is expressed as $O(N)$. For example, if the performance of three conformations is given as $A > B > C$, then one one-membered ensemble (A), one two-membered ensemble (AB), and one three-membered ensemble (ABC) are formed. The fast heuristic results in a small biased sample that neglects the worst performing conformers at each ensemble size.

Each method was implemented in a program called "Ensemble Builder," which was written in the Python language^{58–60} and was used to produce the results reported here. An alpha-version of the Ensemble Builder software is freely available for download through PyPI.

The performance of the three methods was evaluated on the androgen receptor, CDK2, and PPAR- δ . Conformations for these targets were selected from a variety of sources, as summarized in Table 1. Androgen receptor and CDK2

Table 1. Summary of Structures Used To Construct Ensembles

target	structural source	number of structures
androgen receptor	volumetric clustering of five 20 ns MD simulations; one crystal structure	11
CDK2	RMSD clustering of five 20 ns MD simulations; one crystal structure	6
PPAR- δ	wwPDB; Uniprot Q03181	12

structures were selected from pools of five 20 ns MD simulations using two different clustering methods. Volumetric clustering was performed on the binding pocket of the androgen receptor to select 10 conformations, and RMSD clustering was performed on the active site of CDK2, which lead to the selection of five conformations. The crystal structures used to initiate the simulations were also included

Population and Heuristic Samples

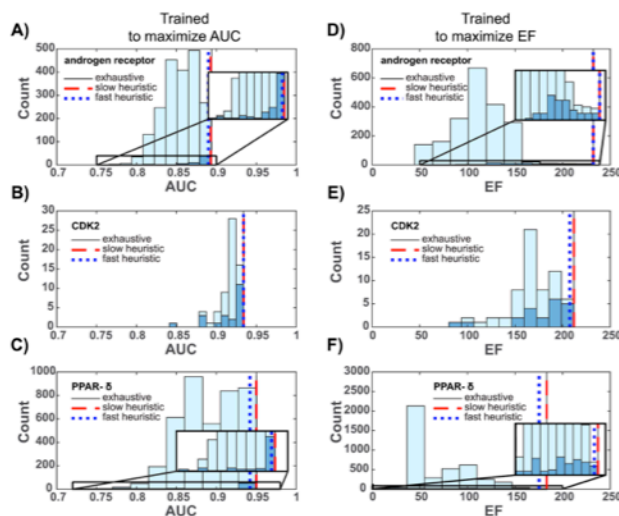


Figure 1. AUC and EF histograms. The exhaustive method was used to enumerate all possible ensembles, and the AUC and EF values of the corresponding population were sorted into 10 bins and plotted in light blue. The slow heuristic method was used to sample a subset of the ensemble population, and the AUC and EF values were sorted into 10 bins and plotted in dark blue. Insets provide expanded views for the androgen receptor and PPAR- δ . AUC values for ensembles trained to maximize the AUC are reported as vertical lines in (A)–(C), and EF values for ensembles trained to maximize the EF are reported as vertical lines in (D)–(F).

and resulted in heterogeneous collections of simulation and experimentally determined conformations of sizes 11 and 6 for the androgen receptor and CDK2, respectively. Twelve human PPAR- δ crystal structures were selected from the protein data bank.⁶¹ Sets of active and decoy compounds for each target were taken from the DUD-E.

The remainder of the Results section is organized as follows. First, the relationship between the ensemble selection algorithms, the anticipated results, and the actual results are examined in the Population and Heuristic Samples section. The dependence of the classification ability on ensemble size is then assessed in the Performance vs Size section, and the results conclude with a comparison of each method on training and test sets in the section entitled Comparing Ensemble Performance on Training and Sets.

Population and Heuristic Samples. Given docking results for an arbitrary collection of target conformers, the exhaustive method enumerates the population of all possible ensembles and identifies the ensemble with the largest objective function value (AUC or EF at a false positive fraction of 0.001). Since the exhaustive method performs a census on the ensemble population and records the performance of each individual, it is guaranteed to identify the best performing ensemble. It follows that if the performance values of the population are represented as a distribution, the value of the best ensemble should reside on the edge of the distribution.

To verify that the best ensemble is found on the edge of the population distribution, ensembles were enumerated, and the

corresponding training set performance values (AUC or EF) were sorted into 10 histogram bins. The resulting distributions are shown in light blue in Figure 1.

Consistent with expectations, the values of the ensembles identified by the exhaustive method appear at the edges of the distributions. This is true across all the targets considered, independent of whether target conformers came from simulation or experiment (Table 1) and provides some confidence in the generality of the approach.

Because the exhaustive method can be computationally expensive, we have developed a more efficient approach, called the “slow heuristic method,” which may have greater utility. To reduce expense, the slow heuristic assumes that the next largest ensemble must contain the current ensemble. Following this assumption, target conformers not yet ensemble members are each grouped with the current best performing ensemble, and the resulting collections are ranked by the values of their objective functions. Hence, the heuristic should result in a population sample biased to favor higher performing ensembles.

To confirm the slow heuristic results in a biased sample that favors higher performance, it was used to construct ensembles, and the corresponding training set AUC and EF values were sorted into 10 bins. The resulting distributions are shown in dark blue in Figure 1.

For each target and both objective functions, the majority of the slow heuristic sample distributions reside on the right side of the corresponding population distributions, which corre-

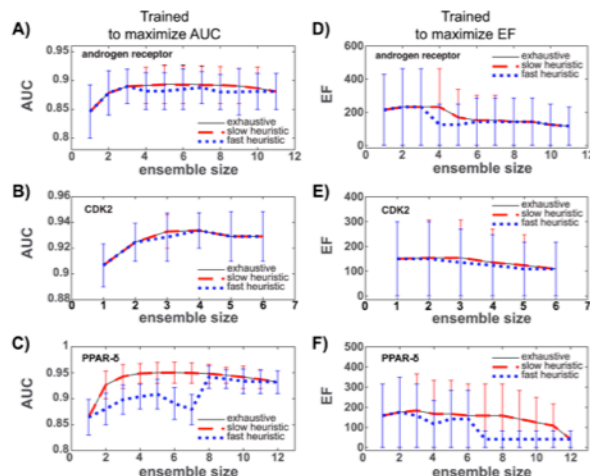


Figure 2. Training set performance as a function of ensemble size for three proteins using DUD-E. AUC is the area under the ROC curve. EF is the ROC enrichment factor at a false positive fraction of 0.001. AUC values for ensembles trained to maximize the AUC are shown in (A)–(C), and EF values for ensembles trained to maximize the EF are reported in (D)–(F). Shown are 95% confidence intervals.

spond to higher performance values. This is consistent with expectations and indicates that recursively generating ensembles from high performing target conformations results in a sample biased to favor performance. The consistency of this result across targets and both AUC and EF values suggests that the approach is generally applicable for a variety targets and ROC-based objective functions.

To further confirm that the slow heuristic produces biased samples, we plotted the performance values of the best ensembles identified by the method as dashed red vertical lines in Figure 1. In five of the six cases considered, the slow heuristic and exhaustive methods result in ensembles that perform identically. In the last case (Figure 1A), the difference was small: an AUC of 0.893 for the slow heuristic compared to a value of 0.894 for the exhaustive method. Since identical results imply that the edges of the samples and populations overlap, these results provide additional evidence that the slow heuristic is able to sample ensembles biased to perform well.

Compared to the population generated by the exhaustive method, the dark blue slow heuristic sample is smaller. The discrepancy between sample and population size becomes larger when a greater number of target conformations is considered. For example, of the three targets, the greatest number of conformations (12) was considered for PPAR- δ , and the difference between the sample and population sizes is largest. This observation is consistent with the scaling of each method: given N conformations, the exhaustive method enumerates a population of size $2^N - 1$, and the slow heuristic method considers samples of size $N(N + 1)/2$.

By assuming that ensembles can be constructed by successively merging target conformations of decreasing performance, the number of ensembles considered is reduced further still, and an approach we call the fast heuristic method is the result. The fast heuristic only considers the performance of

each target conformation once. While this results in the greatest computational efficiency, the method considers the smallest number of ensembles, and the likelihood of failing to sample the best performing ensemble of the population is largest.

Our results indicate that considering a drastically smaller sample of ensembles with the fast heuristic approach does not significantly alter the performance of the best determined ensemble (Figure 1). In four out of six cases, the fast heuristic fails to sample the highest performing ensemble. However, in all cases, the differences in performance are relatively small, and the fast heuristic performance values reside near the edges of the distributions. This indicates that the fast heuristic is able to sample ensembles that perform similarly to the best performing ensemble of the population.

Performance vs Size. When performance is measured as a function of ensemble size (Figure 2), it is notable that for each target the exhaustive method provides an upper bound: this is expected since the exhaustive method identifies ensembles by selecting the top performer from the entire population of a given size.

The slow heuristic and exhaustive methods perform identically, or nearly identically, across the range of ensemble sizes and targets considered. These results are consistent with the distributions shown in Figure 1. However, the trends in Figure 2 go further to imply that the slow heuristic is able to sample the best performing ensemble of the population at each size or an ensemble that performs nearly identically.

While the fast heuristic realizes linear scaling by drastically reducing the number of ensembles considered during training, it is the poorest performing method, particularly for PPAR- δ where the deviations are largest. However, across all targets and for the majority of the sizes considered, the performance values of the fast heuristic fall within the confidence intervals of the exhaustive and slow heuristic methods. Since this implies

Table 2. AUC Values Determined on Training and Test Sets of Best Performing Ensembles Selected To Maximize AUC⁴²

method	androgen receptor			CDK2			PPAR- δ		
	size	training	test	size	training	test	size	training	test
exhaustive	6	0.894 \pm 0.05	0.850 \pm 0.04	4	0.934 \pm 0.014	0.919 \pm 0.019	6	0.950 \pm 0.020	0.923 \pm 0.023
slow heuristic	5	0.893 \pm 0.04	0.850 \pm 0.04	4	0.934 \pm 0.014	0.919 \pm 0.019	6	0.950 \pm 0.020	0.923 \pm 0.02
fast heuristic	3	0.890 \pm 0.03	0.850 \pm 0.04	4	0.934 \pm 0.014	0.919 \pm 0.019	8	0.942 \pm 0.022	0.928 \pm 0.03

⁴²The column labeled "size" gives the number of target conformations in the optimally performing ensemble identified by each method; 95% confidence intervals are given. Androgen receptor ensembles were constructed from 10 MD conformations identified using pocket volume clustering and a crystal structure. CDK2 ensembles were constructed from five MD conformations identified using RMSD-based pocket clustering and a crystal structure. PPAR- δ ensembles were constructed from 12 crystal structures.

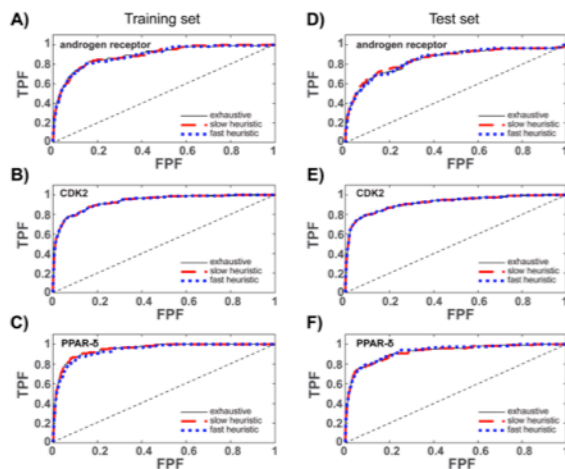


Figure 3. Receiver operating characteristic (ROC) curves for ensembles trained to maximize the AUC of the ROC curve. Dashed black lines illustrate random classification. Training set values are shown in (A)–(C). Test set values are shown in (D)–(F).

performance differences may be attributed to training set variability, the fast heuristic performs reasonably well in comparison to the other two methods.

For a given target and objective function, the smallest and largest ensembles identified by each method perform identically, as the identical bounds of the plots shown in Figure 2 indicate. This behavior is anticipated. Each method forms one-membered ensembles from the single best performing target conformer, and the largest ensembles are formed by merging all target conformations.

Finally, the EF confidence intervals reported in (D)–(F) are larger than those reported for the AUC values in (A)–(C). Because enrichment factors quantify classification performance on a smaller subset of the total data, the larger variability is expected: smaller sample sizes lead to greater standard errors and, by extension, larger confidence intervals.

Comparing Ensemble Performance on Training and Test Sets. When developing knowledge-based classification methods, evaluating the performance of the trained model on an independent test set is a prerequisite to performing a prospective screen. Doing so ensures that the model can correctly classify compounds distinct from the training compounds.³⁴ To further validate the classification ability of the highest performing ensembles identified by each method,

the ensembles were used to screen an independent test set, and the test and training set performances were compared.

As can be seen by examining the androgen receptor entry in Table 2, despite variations in ensemble size and training set performance, the test set results are identical for each method when ensembles are trained using the AUC as an objective function. The variations in ensemble size imply that the samples generated by the slow and fast heuristic do not contain the best performing ensemble of the population. However, the training set performances, which are within confidence intervals of each other, imply that the best performing members of the samples have classification abilities that are similar to the best performing population member. This is consistent with the ROC curves illustrated in Figure 3A and D, which illustrate that the three methods result in ensembles with nearly identical global classification abilities.

Similar results are realized for CDK2, where the three training methods result in identically sized ensembles with identical performance values on both training and test sets; consistently, the ROC curves in Figure 3B and E overlap. Collectively, these results imply that the slow and fast heuristic methods were able to sample the best performing ensemble of the population.

Table 3. EF at FPF of 0.001 Determined on Training and Test Sets of Best Performing Ensembles Selected To Maximize EF at FPF of 0.001^a

method	androgen receptor			CDK2			PPAR- δ		
	size	training	test	size	training	test	size	training	test
exhaustive	4	232.1 \pm 87.2	151.8 \pm 73.9	2	211.9 \pm 57.9	148.3 \pm 50.3	3	183.3 \pm 83.0	116.7 \pm 64.04
slow heuristic	4	232.1 \pm 87.2	151.8 \pm 73.9	2	211.9 \pm 57.9	148.3 \pm 50.3	3	183.3 \pm 83.0	116.7 \pm 64.04
fast heuristic	3	232.1 \pm 87.2	133.9 \pm 70.1	1	207.63 \pm 57.5	152.5 \pm 50.9	2	175.0 \pm 76.6	125.0 \pm 62.2

^aThe column labeled "size" gives the number of target conformations in the optimally performing ensemble identified by each method; 95% confidence intervals are given. Androgen receptor ensembles were constructed from 10 MD conformations identified using pocket volume clustering and a crystal structure. CDK2 ensembles were constructed from five MD conformations identified using RMSD-based pocket clustering and a crystal structure. PPAR δ ensembles were constructed from 12 crystal structures.

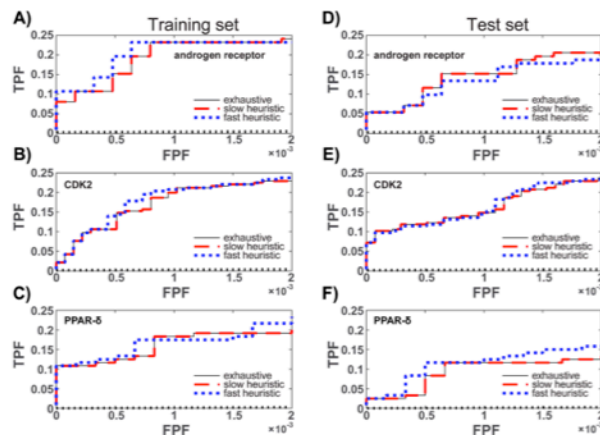


Figure 4. Receiver operating characteristic (ROC) curves for ensembles trained to maximize the EF at a FPF of 0.001. Dotted black lines illustrate random classification. Training set values are shown in (A)–(C). Test set values are shown in (D)–(F). To be consistent with the training condition, the early portion of the ROC curve, with FPF values between 0 and 0.002, is shown.

Consistent with the androgen receptor and CDK2 results, the three methods perform nearly identically on PPAR- δ . By comparing training and test set entries in Table 3, along with Figure 3C and F, it is apparent that the slow heuristic method was able to sample the best performing ensemble from the population, but the fast heuristic method was not. Compared to the best ensemble of the population, the best ensemble sampled by the fast heuristic is slightly larger and performs slightly worse on the training set but slightly better on the test set. However, for both training and test sets, the differences in performance are small, and the AUC values of each method are within confidence intervals of one another. In other words, the best performing ensemble in the fast heuristic sample has similar classification ability as the best performing member of the ensemble population.

Similar to the results produced when using an AUC objective function, each method produces androgen receptor ensembles that perform identically, or nearly so, when ensembles are selected by maximizing the EF. For example, Table 3 shows that the ensembles identified by the exhaustive and slow heuristic methods have identical sizes and performance values. Consistently, Figure 4A and D, which show the early portion of the ROC curves determined on the training and test sets,

respectively, are identical for the exhaustive and slow heuristic methods. While the fast heuristic sample did not contain the optimal population member, the method sampled an ensemble that performed comparably: the performance was identical on the training set and within confidence intervals on the test set.

The pattern is similar when the EF is maximized to identify CDK2 and PPAR- δ ensembles: the slow heuristic samples the best performing member of the population, and the fast heuristic samples an ensemble that performs comparably. In all cases, the performance differences are small, and the averages are within confidence intervals of one another. Collectively, these results provide further evidence that the fast and slow heuristic methods effectively sample ensembles biased to favor high performing members of the population.

Across all the targets considered, the training and test set performances are similar for each method, and similar classification accuracy implies an underlying similarity in the structure of the compounds that make up each set. That is, if training and test set compounds are chemically similar, then they should be classified similarly. To analyze the extent of training and test set overlap, we utilize a popular invariant scaffold representation: graph frameworks.³⁵ A graph framework can be generated from any molecule by converting all

atoms to sp³ hybridized carbon atoms and removing acyclic substructures that do not connect ring systems.

Training and test set similarity was estimated by determining the percentage of molecules whose graph frameworks were unique to each set and the percentage that was shared by each set (Figure 5). Across the three targets, between 65% and 76%

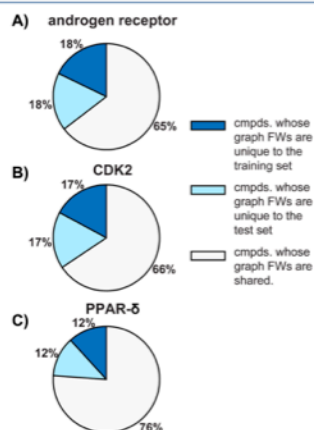


Figure 5. Percentages of compounds whose graph frameworks (FWs) are unique to, and shared between, training and test sets. Percentages are given for (A) androgen receptor, (B) CDK2, and (C) PPAR- δ .

of molecules can be represented by frameworks that are found in both the training and test sets, and between 12% and 18% of molecules are represented by graph frameworks found only in the training or test sets. Hence, the underlying chemical similarities shared by the training and test sets help explain the similar classification performance observed for these sets. However, the existence of molecules whose graph frameworks are unique suggests that the trained ensembles are able to correctly classify molecules structurally distinct from those used during training.

DISCUSSION

Given a collection of target conformations generated either by experiment or by simulation, it is difficult or impossible to know *a priori* which subset will result in the best VS performance. The problem becomes increasingly challenging as the number of target conformations grows, and this is the result of the combinatorial nature of the problem. To address this problem, we presented three knowledge-based ensemble selection methods: the exhaustive method, the slow heuristic method, and the fast heuristic method. For each method, the discussion includes schematic illustrations that describe the underlying selection algorithm and an examination of performance, scaling, and limitations; results from the androgen receptor, CDK2, and PPAR- δ provide context.

Exhaustive Method. By enumeration of all possible combinations of conformers, the exhaustive method generates the complete ensemble population and only retains the highest performing individual; that is, the exhaustive method is

guaranteed to identify the best performing member of the population. This is illustrated schematically in the "Exhaustive" column of Figure 6. Three receptor conformers, colored red,

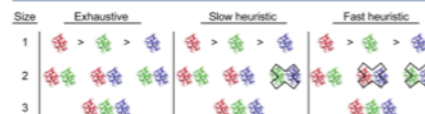


Figure 6. Training method schematic: selecting the best performing ensemble from three target conformers. As indicated by the greater than symbols, the VS performance, measured by either the AUC or EF, is greatest for the red conformer, followed by the green conformer, and the blue conformer is the poorest performer. Each method is found in a column, and all combinatorial possibilities are shown. The VS performance of enumerations marked with an "X" are not explicitly evaluated. Hence, the exhaustive method evaluates all combinatorial possibilities. The slow heuristic assumes the next largest optimal ensemble is formed only from a combination of the current ensemble and one of the remaining conformers. The fast heuristic method ranks the VS performance of each target conformer and assembles ensembles by successively including conformers of decreasing performance.

green, and blue are shown, and the ensembles that can be constructed at each size are also shown. The population constitutes all of the ensembles at each size, and in the simple schematic, contains seven members. By performing a census of the population, the best performing ensemble is readily identified. This was realized for each target considered here.

Applying conformation enumeration to generate ensembles is not new, and the idea has appeared in the literature. For example, to retrospectively compare the VS performance of ensemble and single crystal structure VS protocols, Korb et al.⁶³ used a similar enumerative approach. However, in their work, enumeration was not tied to ensemble training, and we later proposed that enumeration could be used to identify ensembles with the greatest VS utility.⁶³ It is that concept that we demonstrated here.

Slow Heuristic. Performing a population census, as the exhaustive method does, guarantees that the best performing ensemble will be identified, but the process is computationally expensive. To reduce expense, we introduced the slow heuristic, which builds ensembles recursively. Beginning with the best performing target conformer, each conformation not yet assigned to an ensemble is grouped with the best performing ensemble of the current size. This produces a sample of ensembles, each with one additional conformation and a characteristic VS performance, from which the best ensemble is selected. The process continues until all conformers have been included in an ensemble, and the ensemble that performs best overall is retained. Following this heuristic produces a biased sample that neglects population members that do not contain the best performing ensembles of smaller sizes.

To clarify how the slow heuristic results in biased samples, we have illustrated the process schematically in the "Slow heuristic" column in Figure 6. Of the three conformations, the red one performs best, the green next best, and the blue conformation performs worst. The one-membered ensemble is made up of the single best performing target conformer, or the red conformer, in this case. After identifying the one-membered ensemble, two two-membered ensembles are then generated. Each contains the best one-membered ensemble: red-blue and

red-green. Since the blue-green ensemble does not contain the best performing one-membered ensemble, it is neglected.

While it is not a given that the slow heuristic will result in samples biased to favor high performing ensembles, that did prove to be the case in the three targets considered in this work. This was illustrated in part by the overlap of the population and slow heuristic sample distributions in Figure 1 and in part by the ability of the method to identify the best performing ensemble from the population. For example, in Figure 1, the slow heuristic sample favored ensembles that produced larger values of both classification metrics considered, and this was true across all three targets. Additionally, the slow heuristic identified the best performing ensemble from the population in five out of the six cases considered (Tables 2 and 3). These observations provide further support for the claim that the method produces biased samples favoring high performing ensembles, and they suggest that the method may be generally applicable across different target classes and ROC-based objective functions.

Nevertheless, because the slow heuristic samples the population, it may miss ensembles in which synergism between poor performing conformations can lead to a higher performing ensemble. To help clarify this, consider the blue and green conformations in Figure 6. Despite their poorer individual performances, if they pair to form a high performing two-membered ensemble, it will not be sampled by the slow heuristic. However, while missing potential synergism is possible, when the sample is biased toward high performing ensembles, the best performing sample member may perform comparably to the highest performing population member. This proved true in this study. For example, when the slow heuristic was used to train androgen receptor ensembles to maximize the AUC, the sample did not contain the optimal ensemble from the population; however, the performances of the best ensemble from the sample and the best ensemble from the population were within confidence intervals of one another (Table 2).

The slow heuristic appears to offer a reasonable compromise between computational efficiency and performance. To illustrate the computational efficiency, in the Supporting Information, we show that the exhaustive method scales as $O(2^N)$, given N target conformations, while the slow heuristic scales as $O(N^2)$. For example, if 23 receptor conformations are considered, the exhaustive method considers roughly 8.3 million ensembles, while the slow heuristic method only evaluates 264 ensembles. However, since each of the enumerated ensembles can be evaluated on a single processor, it is noteworthy to point out that the exhaustive method is embarrassingly parallel.

Fast Heuristic. By constructing ensembles of increasing size by successively merging conformations of decreasing performance, the fast heuristic ignores the pools of ensembles generated at each size by the slow heuristic and further reduces computational expense. Thus, the fast heuristic produces a small, biased sample that neglects the poorest performing conformations at each ensemble size.

To clarify how the fast heuristic produces biased samples, we have illustrated the process schematically in the "Fast heuristic" column in Figure 6. Since the red conformation performs best, it is selected as the one-membered ensemble, and the poorer performing conformations are neglected. By merging the one-membered ensemble with the next best performing conformation, the two-membered red-green ensemble is produced.

The green-blue ensemble is ignored, just as it is by the slow heuristic, but the red-blue ensemble is also ignored, which results in a smaller sample.

Relative to the exhaustive solution, which generates the entire ensemble population, the fast heuristic sample is significantly smaller. In general, given N target conformations, the ensemble population has a size $2^N - 1$, and the fast heuristic only considers $2N - 1$ of these. In practice, this can quickly amount to thousands of possibilities. For example, with 11 androgen receptor conformations, the fast heuristic method ignores 2026 of the 2047 possible ensembles.

The fast heuristic is nearly identical to the method of Xu and Lill,³⁰ which was described in the Introduction. However, rather than using the value of a ROC classification metric, they ranked target conformations by the differences in average docking scores of decoy and active molecules to each conformer. While Xu's and Lill's results were promising, the effect of ignoring a significant fraction of the ensemble population was not assessed.

To provide insight into the severity of the heuristic relative to the enumerative solution, we compare the exhaustive and fast heuristic results. Consistent with the small sample size, the fast heuristic was only able to identify the best performing population member in one of the six cases considered (Tables 2 and 3). Despite this, the ensembles identified performed similarly to the best performing ensemble of the population: fast heuristic performance values were within confidence intervals of the best performing members of the population for all three targets and both objective functions. Despite the much smaller sample size, the fast heuristic may be a generally applicable approach that offers linear scaling without a dramatic sacrifice in classification ability.

CONCLUSIONS

Docking to structural ensembles is a promising means of identifying novel, structurally diverse active compounds. Despite the potential of ensemble docking, it is difficult to know which structures will synergize and perform well during a virtual screen. This problem emerges from the combinatorial nature of ensemble selection. To address the selection problem, we presented three promising knowledge-based methods. Each method scales differently in the limit of a large number of receptor conformations but can perform similarly, which was demonstrated by constructing ensembles of the androgen receptors, CDK2, and PPAR- δ and with either X-ray crystallographic structures or snapshots from all-atom molecular dynamics trajectories. As with all other knowledge-based methods, those presented here are fundamentally limited by the availability of high quality ligand sets. Nevertheless, virtual screens are often carried out on targets for which active and inactive molecules are known. In these cases, the ensemble selection methods presented have broad applicability.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00684.

Expressions used to determine the standard error of the reported AUC and EF values. Derivations of the big O scaling reported for the exhaustive and slow heuristic methods. (PDF)

DOI: 10.1021/acs.jcim.5b00684
J. Chem. Inf. Model. 2016, 56, 830–842

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ramaro@ucsd.edu.

Author Contributions

S. A. Jusoh and T. L. Offutt contributed equally to the production of this work.

Funding

This work was supported by an NIH Director's New Innovator Award (DP2-OD007237) and an NSF XSEDE Supercomputer resource grant (RAC CHE060073N) to R.E.A. Support from the National Biomedical Computation Resource (NBCR, P41 GM103426) is also gratefully acknowledged. S.A.J. gratefully acknowledges financial support from the Ministry of Education Malaysia and Universiti Teknologi MARA Malaysia.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

R.V.S thanks Sara E. Swift for her thorough reading of the manuscript and substantive discussion and Jesper Sørensen for his useful insights. R.V.S. also thanks Jacob Durrant and Sarah Kochanek for their input.

■ ABBREVIATIONS

ROC, receiver operating characteristic; TPF, true positive fraction; PPF, false positive fraction; AUC, area under the curve; EF, ROC enrichment factor; VS, virtual screening; CDK2, cyclin-dependent kinase 2; PPAR- δ , peroxisome proliferator-activated receptor delta

■ REFERENCES

- Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580–588.
- Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanelo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: a Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867–881.
- Lavecchia, A.; Di Giovanni, C. Virtual Screening Strategies in Drug Discovery: a Critical Review. *Curr. Med. Chem.* **2013**, *20*, 2839–2860.
- Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74–82.
- Schuster, D.; Laggner, C.; Langer, T. Why Drugs Fail – A Study on Side Effects in New Chemical Entities. In *Antitargets: Prediction and Prevention of Drug Side Effects*; Vaz, R. J., Klabunde, T., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2008; pp 3–22.
- Li, A. P. Screening for Human ADME/Tox Drug Properties in Drug Discovery. *Drug Discovery Today* **2001**, *6*, 357–366.
- Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.
- Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.
- McGaughey, G. B.; Sheridan, R. P.; Bayly, C. L.; Culbertson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
- von Korff, M.; Freyss, J.; Sander, T. Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 209–231.
- Amaro, R. E.; Schnauffer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A. Discovery of Drug-Like Inhibitors of an Essential RNA-Editing Ligase in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17278–83.
- Demir, O.; Labaied, M.; Merritt, C.; Stuart, K.; Amaro, R. E. Computer-Aided Discovery of *Trypanosoma brucei* RNA-Editing Terminal Uridyl Transferase 2 Inhibitors. *Chem. Biol. Drug Des.* **2014**, *84*, 131–9.
- Durrant, J. D.; Hall, L.; Swift, R. V.; Landon, M.; Schnauffer, A.; Amaro, R. E. Novel Naphthalene-Based Inhibitors of *Trypanosoma brucei* RNA Editing Ligase 1. *PLoS Neglected Trop. Dis.* **2010**, *4*, e803.
- Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 98–104.
- Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598–1603.
- Swift, R. V.; McCammon, J. A. Substrate Induced Population Shifts and Stochastic Gating in the PBCV-1 mRNA Capping Enzyme. *J. Am. Chem. Soc.* **2009**, *131*, 5126–5133.
- Kumar, S.; Ma, B.; Tsai, C.-J.; Sinha, N.; Nussinov, R. Folding and Binding Cascades: Dynamic Landscapes and Population Shifts. *Protein Sci.* **2000**, *9*, 10–19.
- Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. Molecular Docking to Ensembles of Protein Structures. *J. Mol. Biol.* **1997**, *266*, 424–440.
- Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: an Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.
- Damm, K. L.; Carlson, H. A. Exploring Experimental Sources of Multiple Protein Conformations in Structure-Based Drug Design. *J. Am. Chem. Soc.* **2007**, *129*, 8225–8235.
- Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 198–210.
- Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.
- Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 186–193.
- Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J. F.; Montes, M. Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query. *J. Chem. Inf. Model.* **2013**, *53*, 293–311.
- Bolstad, E. S.; Anderson, A. C. In Pursuit of Virtual Lead Optimization: Pruning Ensembles of Receptor Structures for Increased Efficiency and Accuracy During Docking. *Proteins: Struct., Funct., Genet.* **2009**, *75*, 62–74.
- Nichols, S. E.; Baron, R.; Iveta, A.; McCammon, J. A. Predictive Power of Molecular Dynamics Receptor Structures in Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 1439–1446.
- Ellingson, S. R.; Miao, Y.; Baudry, J.; Smith, J. C. Multi-Conformer Ensemble Docking to Difficult Protein Targets. *J. Phys. Chem. B* **2015**, *119*, 1026–1034.
- Yoon, S.; Welsh, W. J. Identification of a Minimal Subset of Receptor Conformations for Improved Multiple Conformation Docking and Two-step Scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.
- Xu, M.; Lill, M. A. Utilizing Experimental Data for Reducing Ensemble Size in Flexible-Protein Docking. *J. Chem. Inf. Model.* **2012**, *52*, 187–198.
- Katritch, V.; Rueda, M.; Abagyan, R. Ligand-Guided Receptor Optimization. In *Homology Modeling*; Orry, A. J. W., Abagyan, R., Eds.;

- Methods in Molecular Biology; Humana Press: New York, 2012; Vol. 857, pp 189–205.
- (32) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.
- (33) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (34) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.
- (35) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 351–367.
- (36) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations. *J. Mol. Biol.* **2002**, *320*, 597–608.
- (37) Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein pKa Values. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 704–721.
- (38) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of pKa Values for Protein-Ligand Complexes. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 765–783.
- (39) Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.
- (40) Case, D. A.; Berryman, J. T.; Betz, R. M.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Monard, G.; Needham, P.; Nguyen, H.; Nguyen, H. T.; Omelyan, L.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Salomon-Ferrer, R.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; York, D. M.; Kollman, P. A. AMBER 2015; University of California: San Francisco, 2015.
- (41) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
- (42) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (43) Horn, H. W.; Swope, W. C.; Pitner, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.
- (44) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.
- (45) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.
- (46) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (47) Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory Comput.* **2014**, *10*, 5047–5056.
- (48) Daura, X.; van Gunsteren, W. F.; Mark, A. E. Folding-Unfolding Thermodynamics of a β -Heptapeptide from Equilibrium Simulations. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 269–280.
- (49) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (50) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a Software Program for pKa Prediction and Protonation State Generation for Drug-Like Molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.
- (51) LigPrep, version 3.5; Schrödinger LLC: New York, NY, 2015.
- (52) Nicholls, A. Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 1: the Calculation of Confidence Intervals. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 887–918.
- (53) Snedecor, G. W.; Cochran, W. G. The Normal Distribution. *Statistical Methods*, 7; The Iowa State University Press: Ames, IA, 1980; pp 39–63.
- (54) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated Docking to Multiple Target Structures: Incorporation of Protein Mobility and Structural Water Heterogeneity in AutoDock. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 34–40.
- (55) Huang, S. Y.; Zou, X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins: Struct., Funct., Genet.* **2007**, *66*, 399–421.
- (56) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-Dimensional Docking: a Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. *J. Med. Chem.* **2009**, *52*, 397–406.
- (57) Paulsen, J. L.; Anderson, A. C. Scoring Ensembles of Docked Protein-Ligand Interactions for Virtual Lead Optimization. *J. Chem. Inf. Model.* **2009**, *49*, 2813–2819.
- (58) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 10–20.
- (59) van der Walt, S. F.; Colbert, S. C.; Varoquaux, G. I. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.
- (60) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.
- (61) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (62) Korb, O.; Olsson, T. S.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and Limitations of Ensemble Docking. *J. Chem. Inf. Model.* **2012**, *52*, 1262–1274.
- (63) Nichols, S. E.; Swift, R. V.; Amaro, R. E. Rational Prediction with Molecular Dynamics for Hit Identification. *Curr. Top. Med. Chem.* **2012**, *12*, 2002–2012.

Knowledge-Based Methods to Train and Optimize Virtual Screening Ensembles

Robert V. Swift[†], Siti A. Jusoh^{†,§}, Tavina L. Offutt^{†,§}, Eric S. Li[†], Rommie E. Amaro[†]
[†]Department of Chemistry and Biochemistry, University of California San Diego. La Jolla, California, 92093-0340

[‡]Faculty of Pharmacy, Universiti Teknologi MARA Malaysia, 42300 Bandar Puncak Alam, Malaysia

[§]These two authors contributed equally to the production of this work

*Corresponding author: ramaro@ucsd.edu

Supporting Information

Standard Error of the AUC

As given in equation 3 of the paper, the AUC can be expressed by averaging TPF values determined at each inactive compound in the ranked list. Equivalently, the AUC can be determined from an average of the FPF values determined at each active compound in the ranked list; i.e. $AUC = 1 - \langle FPF \rangle_A$. Since the AUC can be determined by averaging over both active and inactive compounds, the numbers of each will contribute to the errors, and each average will have its own distribution. The standard error given in equation 5 of the paper, labeled equation S1 here in the supplementary information, incorporates both error sources.

$$SE = \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_I^2}{N_I}} \quad (S1)$$

The variance due to active and inactive compounds are given as σ_A^2 and σ_I^2 , respectively, while the number of active and inactive compounds included in the estimate are given N_A and N_I , respectively.

The variance due to actives is given by equation S2

$$\sigma_A^2 = \langle (FPF - \langle FPF \rangle_A)^2 \rangle_A \quad (S2)$$

The A subscripts instruct that the averages should be carried out using FPF values determined at each active compound in the ranked list. Similarly, to determine the contribution from the inactive compounds, the following equation is used.

$$\sigma_I^2 = \langle (TPF - \langle TPF \rangle_I)^2 \rangle_I \quad (S3)$$

Standard Error of the ROC Enrichment Factor

The value of the ROC enrichment, equation 4 of the paper, is dependent on the active compounds, through the TPF, and the inactive compounds through the FPF. As a result, error arises from both active and inactive compounds. The standard error can be derived¹ and takes the form giving in equation S4.

$$SE = \frac{1}{FPF} \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_I^2}{N_I}} \quad (S4)$$

Similarly to equation S1, the variances of the active and inactive compounds are given σ_A^2 and σ_I^2 , respectively, while the number of active and inactive compounds included in

the estimate are given N_A and N_I , respectively. The FPF value is the value at which the EF is determined. The variance due to the active compounds is given by equation S5.

$$\sigma^2_A = \frac{1}{FPF^2} \left(\frac{TPF(1-TPF)}{N_A} \right) \quad (S5)$$

Similarly, the variance due to the inactive compounds is given by equation S6

$$\sigma^2_I = \frac{1}{FPF^2} S^2 \left(\frac{FPF(1-FPF)}{N_I} \right) \quad (S6)$$

In equation S6, S^2 is the square of an approximation to the slope of the ROC curve, S , tangent to the point where the EF value was determined. The approximation is derived from an analytic estimate of the ROC curve due to Hanley², $Y = X^{(1-AUC)/AUC}$, as described by Nichols¹.

$$S = EF \left(1 + \frac{\log(EF)}{\log(FPF)} \right) \quad (S8)$$

Training Method Scaling

Exhaustive training. For N conformations, the exhaustive method forms all possible ensembles at each ensemble size from 1 to N . For an ensemble size of k , with $1 < k < N$, the number of ensembles that can be constructed is given by the binomial coefficient,

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} \quad (S8)$$

The total number of ensembles constructed, T , can be determined by summing the values of the binomial coefficient from 1 to N .

$$T = \sum_{k=1}^N \binom{N}{k} \quad (S9)$$

Equation S9 can be simplified by writing the binomial formula as follows.

$$(x + y)^N - x^N = \sum_{k=1}^N \binom{N}{k} x^{N-k} y^k \quad (S10)$$

If we set both x and y to a value of 1 in equation S10 and compare the results to S9, it follows that the total number of constructed ensembles grows exponentially with the number of conformations, as described by equation S11. This growth can be expressed using big O notation as, $O(2^N)$.

$$T = 2^N - 1 \quad (S11)$$

Slow heuristic training. If there are N conformations, in the first step of the slow heuristic method, N one-membered ensembles are considered, and the best performer is retained. In the second step, $N - 1$ two-membered ensembles are considered and the best performer is retained. This process is repeated until a 1 N -membered ensemble is determined. The total number of ensembles constructed, T , is given by equation S12.

$$T = N + N - 1 + N - 2 + \dots + 1 = \sum_{k=1}^N N - (k - 1) \quad (S12)$$

Equation S12 can be re-written as equation S13.

$$T = N(N + 1) - \sum_{k=1}^N k \quad (\text{S13})$$

The sum in the second term of S13 is known as a triangular number and can be re-written as $N(N + 1)/2$, and equation S13 can be simplified.

$$T = \frac{N(N+1)}{2} \quad (\text{S14})$$

In the limit of large N, S14 approaches $N^2/2$. Using big O notation, this is expressed as $O(N^2)$.

References

- (1) Nicholls, A., Confidence Limits, Error Bars and Method Comparison in Molecular Modeling. Part 1: the Calculation of Confidence Intervals. *J Comput.-Aided Mol. Des.* **2014**, *28*, 887-918.
- (2) Hanley, J. A.; McNeil, B. J., The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29-36.

Chapter 4, in full, is a reprint of the material as it appears in Knowledge-Based Methods To Train and Optimize Virtual Screening Ensembles 2015. Swift, Robert V.; Jusoh, Siti A.; Offutt, Tavina L.; Li, Eric S.; Amaro, Rommie E., J Chem Inf Mod, 2016. The dissertation author was a secondary investigator and author of this paper.

Chapter 5

Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor

Ligands

Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands

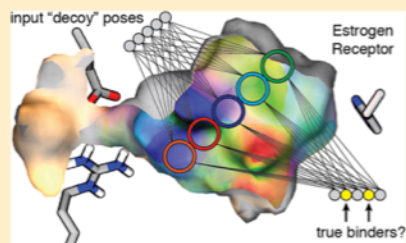
Jacob D. Durrant,[†] Kathryn E. Carlson,[‡] Teresa A. Martin,[‡] Tavina L. Offutt,[†] Christopher G. Mayne,[‡] John A. Katzenellenbogen,[‡] and Rommie E. Amaro^{*†}

[†]Department of Chemistry & Biochemistry and the National Biomedical Computation Resource, University of California, San Diego, La Jolla, California 92093, United States

[‡]Department of Chemistry, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801, United States

S Supporting Information

ABSTRACT: The magnitude of the investment required to bring a drug to the market hinders medical progress, requiring hundreds of millions of dollars and years of research and development. Any innovation that improves the efficiency of the drug-discovery process has the potential to accelerate the delivery of new treatments to countless patients in need. “Virtual screening,” wherein molecules are first tested in silico in order to prioritize compounds for subsequent experimental testing, is one such innovation. Although the traditional scoring functions used in virtual screens have proven useful, improved accuracy requires novel approaches. In the current work, we use the estrogen receptor to demonstrate that neural networks are adept at identifying structurally novel small molecules that bind to a selected drug target, ultimately allowing experimentalists to test fewer compounds in the earliest stages of lead identification while obtaining higher hit rates. We describe 39 novel estrogen-receptor ligands identified in silico with experimentally determined K_i values ranging from 460 nM to 20 μ M, presented here for the first time.



INTRODUCTION

There is an urgent need for innovative approaches to improve the efficiency of the drug-discovery process. The purpose of the current work is to highlight the potential benefits of applying machine learning, specifically neural networks, to structure-based drug discovery. Artificial neural networks (ANN), first conceived in the 1950s,⁵ have become popular in recent decades thanks to algorithmic and hardware advances. Although ANN have been applied to drug discovery in the context of ligand-based QSAR (see, for example, ref 6), they have not traditionally been used in structure-based virtual-screening methods. We here confirm that they are well suited to this important task. Given the ever-growing amount of data available for training^{7–9} and the recent evolution of GPU-accelerated computation, we believe neural-network-based techniques have the potential to transform the in silico prediction of molecular recognition.

High-throughput biochemical screens are often used to identify pharmacologically active compounds. Although highly automated, these screens require specialized hardware, labor, and carefully managed consumables, making them nontrivial and cost-intensive endeavors that are inaccessible to many researchers in academia and industry. In silico techniques such as virtual screening require only modest computational

infrastructure and have become an attractive alternative for lead identification.

Structure-based virtual screening is a two-step process in which a molecule is first docked (i.e., positioned) into a receptor pocket and then evaluated using a scoring function that predicts activity. Reliable scoring functions are required to effectively enrich a set of top-predicted binders with potential hits.^{10–16} Great effort has been dedicated to improving their accuracy, although much room for improvement remains.

Durrant et al. recently created two fast and accurate neural-network scoring functions for rescoring docked ligand poses (NNScore 1.0 and 2.0).^{17–19} Unlike traditional docking scoring functions, these nonparametric functions are not constrained to predetermined physical formulas or statistical analyses; rather, they “learn” directly from existing experimental data how best to predict binding and so can, in theory, better capture the nonlinear, synergistic relationships among binding determinants. To our knowledge, these are the first neural-network scoring functions that predict affinity by directly examining atomic-resolution ligand–protein interactions.

Machine-learning docking rescoring functions in general, and NNScore in particular, have only recently been described in the

Received: April 27, 2015
Published: August 18, 2015

literature. Initial studies have shown that this class of scoring functions performs well in *retrospective* studies, as judged by the ability to predict previously determined experimental binding affinities²⁰ or to separate known ligands from a larger library of presumed nonbinding decoy molecules.¹⁷ However, with some notable exceptions (see, for example, refs 21–23), these kinds of functions have not been extensively used to *prospectively* identify novel ligands, as required for drug discovery.

The purpose of the current work is to provide additional evidence that NNScore is in fact well suited to prospective drug discovery. Building on one of our previous studies,¹⁷ we here use NNScore to identify 39 novel ligands of the estrogen receptor (ER), the target of several drugs used clinically to treat breast cancer,^{24,25} osteoporosis,²⁴ anovulation,²⁶ dyspareunia,²⁷ and male hypogonadism.²⁸

RESULTS AND DISCUSSION

Background: Neural Networks. The NNScore scoring function is based on artificial neural networks, machine-learning modules that are designed to mimic, albeit inadequately, the microscopic architecture of the brain. Virtual neurons, called neurodes, are connected by virtual axons, called connections. In brief, information to be analyzed is encoded on a set of neurodes called the input layer. This information is processed as it cascades through the neurodes of the network. The final analysis is encoded on a group of neurodes called the output layer. Neural networks are trained by gradually adjusting the connection strengths until the networks can reliably predict the correct output from a given input.

In previous studies, we trained neural networks to predict small-molecule/receptor binding by first generating numeric “descriptors” of thousands of crystallographic binding poses.^{18,19} The descriptors used to train NNScore 1.0 included tallies and categorizations of juxtaposed ligand/receptor atoms, summed electrostatic energies, ligand atom types, and rotatable-bonds counts. Training NNScore 2.0 similarly relied on tallies and categorizations of juxtaposed ligand/receptor atoms and summed electrostatic energies, as well as (1) additional molecular interactions/properties as determined by the BINANA algorithm²⁹ and (2) physics-based terms borrowed from the AutoDock Vina scoring function.³⁰

Neural networks were trained to predict the strength of binding from these descriptors by fitting against experimentally measured binding affinities. Specifically, NNScore 1.0 was trained to categorize ligands by potency (high-affinity vs low-affinity binder). In contrast, NNScore 2.0 was trained to predict the binding affinity directly.

Following this training phase, other small-molecule/receptor binding poses to which the networks had never been previously exposed (e.g., from docking studies) could be similarly analyzed. For a given set of binding-pose descriptors, the networks return a score that correlates with the likelihood of high-affinity binding. When a list of docked compounds is ordered by this score, the set of top-ranked molecules is often enriched for true ligands.

In a recent study, we compared the retrospective virtual-screening performance of NNScore 1.0 and 2.0 across ~40 diverse protein receptors (Figure 1A). This benchmark study suggested that the average performance of NNScore 1.0 is better than that of NNScore 2.0. However, NNScore 2.0 was the superior function for some receptors,¹⁷ highlighting the utility of employing multiple scoring functions in any computer-aided drug-discovery (CADD) project.

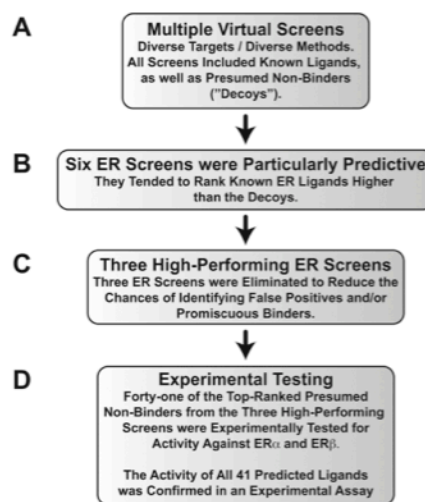


Figure 1. Computational/experimental protocol used to identify novel estrogen-receptor ligands.

Background: NNScore Performance against the Estrogen Receptor. The estrogen receptor alpha (ER α) was among the ~40 diverse receptors considered previously.¹⁷ Both ER α and the highly homologous ER β , which differ by only two binding-pocket amino acids³¹ (56% sequence identity in the ligand-binding domain³²), are attractive drug targets.^{32,33} These transcription factors are activated by the endogenous steroid hormone 17 β -estradiol, leading to gene regulation via binding to specific DNA target sequences. A number of ER α ligands, many of which are nonsteroidal, are currently FDA approved for the treatment of osteoporosis,²⁴ breast cancer,^{24,25} anovulation,²⁶ dyspareunia,²⁷ and male hypogonadism.²⁸ ER β is emerging as a promising cancer, cardiovascular, inflammatory, and central-nervous-system drug target.^{33,34} Various small molecules, including some approved drugs, act as agonists, antagonists, or mixed-function partial agonist/antagonists. The level of agonist vs antagonist activity depends on binding-induced ER-receptor conformational changes,^{2,3} as well as on the cellular and even tissue context.³⁵

In the previous retrospective virtual-screening study, we used a small-molecule library consisting of known ER α ligands and presumed decoys (e.g., molecules presumed to be nonbinders for testing purposes, though without experimental conformation) taken from the Directory of Useful Decoys (DUD)³⁶ and the NCI diversity set III (<http://dtp.nci.nih.gov/>), respectively. The DUD-set ER α agonists and antagonists were docked into their respective ER α structures in the agonist- or antagonist-bound conformations, as appropriate; the same set of NCI decoys was used for both receptors. For each conformation, seven distinct docking/scoring protocols involving AutoDock Vina,³⁰ Schrödinger's Glide,³⁷ and NNScore^{18,19} were employed.

In six of these virtual screens, over ~75% of the known ligands were contained in the set of top-ranking compounds

large enough to include 5% of the presumed decoys (i.e., the true positive rate was $> \sim 75\%$ when the false-positive rate was fixed at 5%, Figures 1B and 2). This metric, which we call the “metric of early performance,”¹⁷ indicates how well a given scoring function is able to enrich the top-ranking compounds with true ligands.

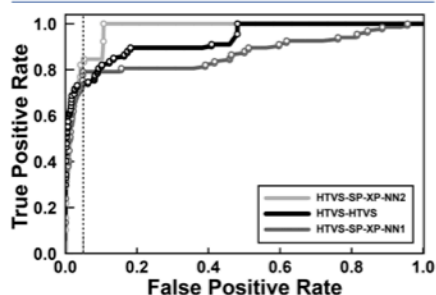


Figure 2. ROC curves associated with each of the three high-performing virtual screens. The data points corresponding to the known ER ligands are shown as circles. The vertical dotted line corresponds to a false positive rate of 5%, used to calculate the early performance metric.

While the vast majority of the top-ranking “decoy” molecules used in the initial study were certainly not true ER ligands, in the current study we hypothesize that some might in fact be true binders. By showing that this hypothesis is correct, we provide evidence that NNScore can prospectively identify novel ligands from among decoys and therefore has potential for use in structure-based computer-aided drug discovery.

Ignoring Potentially Promiscuous Top-Ranked Compounds. In previous virtual-screening studies, we have noted that certain molecules have a tendency to frequently appear among the top-ranked compounds, even when targeting diverse and unrelated receptors. There are two possible explanations for this phenomenon. First, these compounds may in fact bind to many diverse targets, in which case they are promiscuous and so are poor candidates for drug discovery. Second, they may in fact be nonbinders (false positives) that the scoring functions incorrectly identify as ligands due to inappropriate biases. In either case, such compounds are arguably not worth pursuing.

The six high-performing ER virtual screens described above identified a number of potentially promiscuous and/or false-positive compounds. Several of the top-ranked compounds were also frequently present among the top-ranked compounds of other high-performing screens from the previously published retrospective study, even though those screens targeted unrelated receptors. To enhance our chances of identifying true and useful ER ligands, we therefore discarded all virtual “hits” that were found among the top compounds in more than three of the high-performing virtual screens (Figure 1C).

This filtering process had a substantial impact on three of the six high-performing ER virtual screens. Of the top-ranked compounds from these screens, 14/15 or 15/15 were judged problematic. Given that our goal was to ultimately submit only the top compounds from each screen for experimental testing, we opted to focus exclusively on the other three high-

performing ER screens that were less affected. These screens used the following protocols: (1) compounds were docked with a three-tiered Glide protocol (HTVS/SP/XP) into the ER antagonist conformation and then rescored with NNScore 2.0 (HTVS-SP-XP-NN2/Antagonist); (2) compounds were docked and scored with Glide HTVS into the ER agonist conformation (HTVS/Agonist); and (3) compounds were docked with a three-tiered Glide protocol into the ER agonist conformation and then rescored with NNScore 1.0 (HTVS-SP-XP-NN1/Agonist).

It was fortunate that virtual screens against both the agonist- and antagonist-bound ER structures performed well. When a small molecule approaches its receptor *in vivo*, it encounters a flexible binding pocket in constant motion, not a single crystalline conformation.³⁸ This is especially true of the highly dynamic estrogen-receptor binding pocket, which can assume different geometries depending on the size and shape of the bound ligand.^{39,40} Even a scoring function with perfect accuracy could not identify ligands that bind to pockets with unconsidered geometries. By including multiple structurally diverse receptor conformations in virtual-screening campaigns, ligands with a broader diversity of binding poses can potentially be identified.⁴¹

Evidence for the predictive utility of these three virtual screens was apparent even prior to experimental testing, as two known ER ligands inadvertently included among the 1560 presumed decoys were correctly identified (genistein, identified by HTVS, and naringenin, identified by both HTVS and HTVS-SP-XP-NN1).

Experimental Confirmation. Forty-one compounds, including genistein and naringenin, were tested experimentally for ER binding using a competitive radiometric ligand binding assay with an operational sensitivity (limit of detection) of $K_i < 20 \mu\text{M}$ (Figure 1D).⁴² Remarkably, all molecules predicted to be ER ligands *in silico* had experimental K_i values less than $8 \mu\text{M}$. Excluding genistein and naringenin, the most potent novel ER ligands were NCI-19136, NCI-33005, and NCI-13151, with K_i values of 460, 780, and 1380 nM, respectively (Table 1). Each of these compounds was coincidentally found using a different docking protocol, suggesting that applying multiple CADD techniques to a given target can also increase the diversity of the identified ligands.⁴³ Though the virtual screens targeted ER, a similar experimental assay revealed that all 41 compounds bound to ER β as well (K_i values $\leq 20 \mu\text{M}$). NCI-33005, NCI-13151, and NCI-19136 were notable ER β binders, with K_i values of 330, 1540, and 2000 nM, respectively (Figure 1D).

These results suggest that (1) NNScore is well suited to prospective drug-discovery projects targeting this system and (2) NNScore can complement more classical scoring functions.

Comparison of Docking Methods. Twenty-nine of the 39 novel ligands presented here for the first time were initially identified using one of the two NNScore protocols, 15 were identified using HTVS, and 3 were identified by both methods. The average K_i values of the compounds found using the HTVS-SP-XP-NN2/Antagonist, HTVS/Agonist, and HTVS-SP-XP-NN1/Agonist protocols were 4.12, 3.68, and 4.10 μM , respectively. A one-way ANOVA analysis led us to reject the null hypothesis that these average K_i values were statistically different ($p = 0.76$), suggesting that the three protocols performed comparably.

Multiple studies have demonstrated that scoring functions are remarkably receptor specific (see, for example, refs 17 and

Table 1. High-Affinity Compounds Found by Docking into ER α Structures in Both the Antagonist- and Agonist-Bound Conformations, Sorted by the Experimentally Measured ER α K $_i$ ^{3,4}

Compound	Structure	ER α K $_i$ (μ M)	ER β K $_i$ (μ M)	HTVS-SP-XP-NN2 Percentile	HTVS Percentile	HTVS-SP-XP-NN1 Percentile
NCI-19136 (Figure 4A)		0.46 \pm 0.004	2.00 \pm 0.4	3.38	(9.47)	(>12.10)
NCI-33005 (Figure 4B)		0.78 \pm 0.2	0.33 \pm 0.1	(5.07)	2.95	(10.88)
NCI-36586 (Genistein)		0.79 \pm 0.1	0.008 \pm 0.00008	(9.69)	2.27	(5.10)
NCI-13151 (Figure 4C)		1.38 \pm 0.3	1.54 \pm 0.3	(>12.50)	(25.88)	2.58
NCI-308849		1.38 \pm 0.1	4.83 \pm 0.1	(>12.50)	(10.76)	2.27
NCI-17128		1.81 \pm 0.4	4.91 \pm 1.4	(>12.50)	(9.40)	3.20
NCI-122253		1.98 \pm 0.2	5.29 \pm 0.9	1.13	1.41	(5.78)
NCI-130847		2.05 \pm 0.2	6.75 \pm 1.5	2.19	(33.74)	(>12.10)
NCI-165701		2.47 \pm 0	3.16 \pm 0.9	(>12.50)	3.38	(5.96)
NCI-34875 (Naringenin)		2.56 \pm 0.5	3.10 \pm 0.9	(4.94)	2.83	3.93
NCI-78623		2.78 \pm 0.8	3.99 \pm 0.6	(11.76)	1.66	(11.25)
NCI-351674		2.82 \pm 0.6	8.25 \pm 1.4	(5.13)	3.69	(7.68)

Table 1. continued

Compound	Structure	ER α K _i (μ M)	ER β K _i (μ M)	HTVS-SP-XP-NN2 Percentile	HTVS Percentile	HTVS-SP-XP-NN1 Percentile
NCI-12262		3.08 \pm 1.0	11.90 \pm 0	(>12.50)	(37.86)	3.44
NCI-201863		3.18 \pm 0.2	7.00 \pm 0.4	1.38	(>49.90)	(>12.10)
NCI-95909		3.74 \pm 0.9	7.48 \pm 1.4	(>12.50)	(5.10)	1.97
NCI-112541		4.14 \pm 0.9	5.26 \pm 1.3	2.63	(15.49)	(4.79)
NCI-246999		4.32 \pm 1.0	3.78 \pm 0.1	3.13	(>49.90)	(>12.10)
NCI-319709		5.20 \pm 1.2	6.88 \pm 0.7	2.94	(>49.90)	(>12.10)
NCI-117554		5.25 \pm 1.1	7.48 \pm 1.4	(5.63)	(5.47)	3.38
NCI-111847		5.42 \pm 0.4	7.43 \pm 0.5	1.63	(4.12)	3.87

^aAdditional experimentally validated ligands are described in the Supporting Information. Note that the compounds themselves were tested only for binding, not for agonism vs antagonism. For each docking protocol/compound, we report the percentile rank (NCI and DUD compounds considered together). When a given compound did not rank high enough to warrant experimental follow up, the percentile is given in parentheses. Additionally, a lower bound on the percentile is given for compounds that could not be docked/scored at all. Additional compounds are listed in Tables S1, S2, and S3.

44). Similarly, scoring-function performance may be affected by certain chemical features of the small molecules being screened, especially features used to train the scoring functions themselves. One crude way of measuring potential ligand-based biases is to assess the structural diversity of the ligands identified in a given virtual screen. While scoring functions are not typically trained to maximize ligand diversity, functions that identify a set of validated ligands with substantially reduced structural diversity relative to the source library are perhaps suspect.

While hit diversity can be a useful performance metric, we wish to emphasize its limitations. Screens with low hit diversity are not necessarily flawed. One might expect a virtual screen to pull out clusters of structurally analogous true ligands. Similarly, screens with high hit diversity are not necessarily free of bias. A

scoring function that inappropriately favors compounds with chemical properties that are structure independent (e.g., high molecular weight, high hydrophobicity, etc.) could incorrectly identify false-positive "hits" that are nonetheless structurally diverse. But we do believe that a substantial lack of chemical diversity between compound clusters may in some cases indicate that the associated scoring functions have been over fitted to favor the known, explored, and nonprotectable chemotypes that generally comprise scoring-function training sets. Generally speaking, an ideal virtual screen should identify diverse and unique molecules, in addition to identifying high-affinity compounds.

Compound diversity and uniqueness can be assessed by classifying compounds according to molecular scaffolds (e.g., molecular graphs).^{45–49} The NCI Diversity Set III (NCIDSIII),

Table 2. Chemical-Diversity Analysis Using Molecular Graphs^{a†}

compound set	N [number]	N_G [molecular graphs]	N_G/N [diversity ratio]	N_s [singleton graphs]	N_s/N [singleton ratio]
NCI Diversity Set III	1560	652	0.42	475	0.31
HTVS-SP-XP-NN2/Antagonist	15	15	1.0	15	1.0
HTVS/Agonist	15	14	0.93	13	0.87
HTVS-SP-XP-NN1/Agonist	15	13	0.87	12	0.80

^aN = total number of compounds in the library, N_G = number of molecular graphs in the library, N_G/N = diversity ratio, N_s = number of singleton molecular graph scaffolds, N_s/N = singleton ratio.

which contained the structurally diverse presumed decoys used in the retrospective virtual screens, spanned 652 molecular graphs (Table 2). To facilitate subsequent comparison with other compound sets, this count was normalized by the total number of library compounds, giving a unique-framework (diversity) ratio (i.e., structurally distinct scaffolds_{count}/total number of compounds_{count}) of 0.42. To put this number into perspective, if each library compound had a unique graph (i.e., optimal diversity), this ratio would be 1.0. In contrast, if all compounds were analogs with the same scaffold (minimal diversity), the ratio would be close to zero.

As a complementary metric, we also measured the uniqueness of the NCI compounds. 475 molecular graphs were associated with a single compound (i.e., singletons, Table 2). A NCIDSIII singleton ratio of 0.31 was calculated by dividing the number of singletons by the total number of compounds.

We next measured the diversity of the three sets of hits identified using the HTVS-SP-XP-NN2/Antagonist, HTVS/Agonist, and HTVS-SP-XP-NN1/Agonist virtual-screening protocols, respectively. The top hits found using these three methods were comparably diverse. The diversity ratios were 1.0, 0.93, and 0.87, respectively (Table 2). Similarly, the singleton ratios were 1.0, 0.87, and 0.80 for the HTVS-SP-XP-NN2/Antagonist, HTVS/Agonist, and HTVS-SP-XP-NN1/Agonist hits, respectively (Table 2). In all cases, the hits were judged to be even more diverse and enriched in singletons than the NCIDSIII compounds generally.

The fact that our top hits were more diverse and unique than the broader NCIDSIII suggests that the docking protocols used do not unduly favor certain molecular scaffolds.

Binding Poses. The BINANA algorithm²⁹ was used to identify potential receptor–ligand interactions between the crystallographic pose of estradiol (Figure 3), the native ligand, and the docked poses of NCI-19136, NCI-33005, and NCI-13151 (Figure 4), the three highest affinity novel ER α ligands identified. NCI-33005 and NCI-13151 had very similar docked poses, as did NCI-19136 when the 2H-pyrazole tautomer was considered. Like the native ligand estradiol, NCI-19136 and NCI-33005 are predicted to form hydrogen bonds with residue E353. In contrast, the NCI-13151 pose forms a hydrogen bond with the L387 backbone carbonyl oxygen atom, though a simple rotation of the NCI-13151 hydroxyl group would easily permit a hydrogen bond with E353. Like estradiol, all three ligands may also form T-shaped π – π interactions with F404. NCI-33005 and NCI-13151 may also form hydrogen bonds with R394, just as the estradiol phenyl hydroxyl group does.

In other ways, the three novel ligands have predicted binding poses unlike that of estradiol. For example, NCI-19136 and NCI-33005 may form additional hydrogen bonds with the L346 backbone carbonyl oxygen atom, and NCI-13151 may form a hydrogen bond with L387, as mentioned above. Also, neither of the novel ligands appears to form a hydrogen bond with HS24, apparently failing to exploit one of the interactions

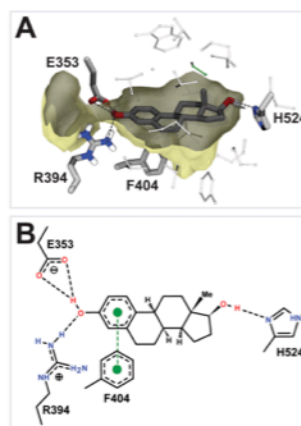


Figure 3. Crystallographic pose of estradiol. (A) The 1G50 crystal structure.¹ (B) Schematic of estradiol binding, modified from output generated by PoseView.⁴

characteristic of estradiol binding. Both ligands do extend aromatic moieties in the direction of HS24, however. Modifying these moieties so they can approach and interact with HS24 may be an effective drug-optimization strategy, though we do note that a number of other ER ligands (e.g., afimoxifene² and raloxifene³⁰) have high binding affinities even in the absence of this interaction.

Conclusion. Although the novel scaffolds presented here may be useful for future drug development, ER α has been extensively studied and is already the target of several highly optimized FDA-approved drugs (e.g., tamoxifen, fulvestrant, and raloxifene). The primary utility of the current work is therefore to demonstrate the remarkable performance of two recently developed neural-network docking rescoring functions,^{18,19} which, according to a recent retrospective study, are effective against this target as well as a number of others.¹⁷ Herein, we have shown by further computational and experimental analyses that several high-scoring presumed “decoys” indeed bind the receptor target with low micromolar affinity, indicating that these scoring functions are effective when employed prospectively.

Nonparametric machine-learning techniques such as neural networks are often used in ligand-based QSAR, but their application as receptor-centric docking rescoring functions is less common.^{18–20,51} These scoring functions take a novel approach to predicting molecular recognition. We believe they

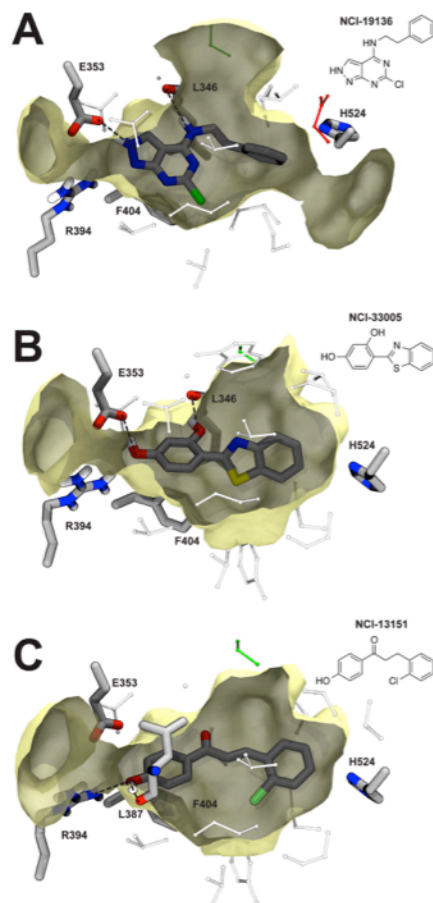


Figure 4. Binding poses. (A) NCI-19136 docked into the 3ERT antagonist structure. Dotted black lines represent hydrogen bonds. (B) NCI-33005, docked into the 1L2I agonist structure.⁹ (C) NCI-13151, docked into the 1L2I agonist structure. A potential hydrogen bond with E353 is less certain and so is not shown.

can complement more traditional scoring functions, helping to identify ligand scaffolds that might not be found otherwise.

MATERIALS AND METHODS

Computational Details. Durrant et al. performed a retrospective virtual-screening benchmark study in 2013 to assess the performance of two novel neural-network-based scoring functions, NNScore 1.0 and NNScore 2.0.^{17–19} Human estrogen receptor alpha (ER α) in both the agonist- (ER α) and antagonist-bound (ER α) conformations were among

the ~40 diverse receptors considered. In brief, models of ER α agonist and ER α antagonist were prepared from published crystal structures (PDB IDs 1L2I⁹ and 3ERT,² respectively). Molecular models of known ER α agonists and antagonists obtained from the Directory of Useful Decoys (DUD)²⁰ were docked into the relevant receptor, together with 1560 diverse small molecules from the NCI diversity set III (<http://dtp.nci.nih.gov/>) that served as presumed decoys.

Two high-performing ER α virtual screens targeting ER α agonist and ER α antagonist, respectively, used a multistep docking protocol. Compounds were first docked into each receptor using Schrödinger's Glide HTVS, a fast program designed for high-throughput virtual screening. The compounds were then ranked by the docking score, and the top 50% were subsequently docked using Glide SP, a more computationally demanding program thought to be more accurate. The top 50% of the Glide-SP-docked compounds were then docked using Glide XP, Schrödinger's most rigorous program. Finally, the XP ER α agonist and ER α antagonist poses were rescored using NNScore 1.0¹⁸ and NNScore 2.0,¹⁹ respectively. These two docking protocols are here called HTVS-SP-XP-NN1/Agonist and HTVS-SP-XP-NN2/Antagonist, respectively. Durrant et al. also obtained an early performance metric¹⁷ of 79% when Glide HTVS alone was used to dock compounds into ER α agonist (HTVS/Agonist).

The early performance metric used to assess these three virtual screens was predicated on the assumption that the NCI compounds do not in fact bind to ER α . This assumption is certainly true for the vast majority of these structurally diverse compounds, but it is likely that at least some of the NCI compounds are in fact true ER α ligands. We therefore selected the 15 top-ranked, nonpromiscuous NCI compounds from each of the three high-performing ER α virtual screens for subsequent experimental testing. As some compounds ranked well in multiple screens, forty-one molecules were advanced in total.

Experimental Details. Relative binding affinities were determined by a competitive radiometric binding assay with 2 nM [³H]estradiol as tracer (PerkinElmer, Waltham, MA), as described previously.⁵² Full-length purified human ER α and ER β were purchased from Pan Vera/Invitrogen (Carlsbad, CA). Following incubation for 18–24 h at 0 °C, the receptor–ligand complexes were absorbed onto hydroxyapatite (BioRad, Hercules, CA) and unbound ligand was washed away.⁵² All small molecules tested were taken from the NCI Diversity Set III and have purities over 90% per LC/Mass Spec.

Affinities were initially expressed as relative binding affinity (RBA) values, where the RBA of estradiol is set at 100%. Under these conditions, the K_d values of estradiol are ~0.2 and ~0.5 nM for ER α and ER β , respectively. These RBA measurements were reproducible in separate experiments with coefficients of variation of 0.3. The values shown in Table S4 represent the average plus or minus the standard deviation, calculated using two or more independent measurements. The K_i values reported in Table 1 were calculated by dividing the average K_d of estradiol by the RBA and then multiplying by 100.⁵³

Chemical Diversity of the Hits. To analyze the chemical diversity and uniqueness of the NCI Diversity Set III, as well as the novel hits identified using the HTVS-SP-XP-NN2/Antagonist, HTVS/Agonist, and HTVS-SP-XP-NN1/Agonist docking protocols, we considered two distinct scaffold representations: Bemis-Murcko frameworks (see the Supporting Information) and molecular graphs (described in the main

text). Bemis-Murcko frameworks consist of any ring system and linker groups,⁴⁸ and molecular graphs consist of nodes (carbon atoms) connected via edges (single bonds). Unlike Bemis-Murcko frameworks, molecular graphs exclude any atom-type or bond-order information. Both the frameworks and graphs were generated using the RDKit package MurckoScaffold.⁵⁴

■ ASSOCIATED CONTENT

5 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00241.

The Supporting Information lists additional experimentally validated ER ligands beyond those found in Table 1. Further experimental results and analyses of molecular diversity/uniqueness are also provided. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ramaro@ucsd.edu.

Funding

This work was supported by an NIH Director's New Innovator Award (DP2-OD007237) and an NSF XSEDE Supercomputer resources grant (RAC CHE060073N) to R.E.A., as well as an NIH grant (DK015556) to J.A.K. and an NIH training fellowship (T32ES007326) to T.A.M. Support from the National Biomedical Computation Resource (NBCR, P41 GM103426) is also gratefully acknowledged.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Aaron J. Friedman for performing the original Glide virtual screens.

■ ABBREVIATIONS

ER, estrogen receptor; ER α , estrogen receptor alpha; ER β , estrogen receptor beta; NCIDSIII, National Cancer Institute's Diversity Set III; ER_{agonist}, estrogen receptor in the agonist-bound conformation; ER_{antagonist}, estrogen receptor in the antagonist-bound conformation; DUD, Directory of Useful Decoys; RBA, relative binding affinity

■ REFERENCES

- (1) Eiler, S.; Gangloff, M.; Duclaud, S.; Moras, D.; Ruff, M. Overexpression, Purification, and Crystal Structure of Native Er Alpha Lbd. *Protein Expression Purif.* **2001**, *22*, 165–173.
- (2) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell* **1998**, *95*, 927–937.
- (3) Shiau, A. K.; Barstad, D.; Radek, J. T.; Meyers, M. J.; Nettles, K. W.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A.; Agard, D. A.; Greene, G. L. Structural Characterization of a Subtype-Selective Ligand Reveals a Novel Mode of Estrogen Receptor Antagonism. *Nat. Struct. Biol.* **2002**, *9*, 359–364.
- (4) Stierand, K.; Maass, P. C.; Rarey, M. Molecular Complexes at a Glance: Automated Generation of Two-Dimensional Complex Diagrams. *Bioinformatics* **2006**, *22*, 1710–1716.
- (5) Rosenblatt, F. The Perceptron - a Probabilistic Model for Information-Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65*, 386–408.

- (6) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in Qsar and Qspr. *J. Chem. Inf. Model.* **2002**, *42*, 903–911.
- (7) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The Pdbbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.
- (8) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The Pdbbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.
- (9) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding Moad (Mother of All Databases). *Proteins: Struct., Funct., Genet.* **2005**, *60*, 333–340.
- (10) Huang, S. Y.; Grinter, S. Z.; Zou, X. Q. Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899–12908.
- (11) Smith, R. D.; Dunbar, J. B.; Ung, P. M. U.; Esposito, E. X.; Yang, C. Y.; Wang, S. M.; Carlson, H. A. Csar Benchmark Exercise of 2010: Combined Evaluation across All Submitted Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115–2131.
- (12) Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on Pdbbind Database. *J. Comput. Chem.* **2011**, *32*, 742–755.
- (13) Wang, J. C.; Lin, J. H. Scoring Functions for Prediction of Protein-Ligand Interactions. *Curr. Pharm. Des.* **2013**, *19*, 2174–2182.
- (14) Cheng, T. J.; Li, Q. L.; Zhou, Z. G.; Wang, Y. L.; Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: A Problem-Centric Review. *AAPS J.* **2012**, *14*, 133–141.
- (15) Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and Advances in Computational Docking: 2009 in Review. *J. Mol. Recognit.* **2011**, *24*, 149–164.
- (16) Meng, X. Y.; Zhang, H. X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 146–157.
- (17) Durrant, J. D.; Friedman, A. J.; Rogers, K. E.; McCammon, J. A. Comparing Neural-Network Scoring Functions and the State of the Art: Applications to Common Library Screening. *J. Chem. Inf. Model.* **2013**, *53*, 1726–1735.
- (18) Durrant, J. D.; McCammon, J. A. Nnscore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865–1871.
- (19) Durrant, J. D.; McCammon, J. A. Nnscore 2.0: A Neural-Network Receptor-Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51*, 2897–2903.
- (20) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169–1175.
- (21) Lindert, S.; Zhu, W.; Liu, Y. L.; Pang, R.; Oldfield, E.; McCammon, J. A. Farnesyl Diphosphate Synthase Inhibitors from in Silico Screening. *Chem. Biol. Drug Des.* **2013**, *81*, 742–748.
- (22) Zhu, W.; Zhang, Y.; Sinko, W.; Hensler, M. E.; Olson, J.; Molohon, K. J.; Lindert, S.; Cao, R.; Li, K.; Wang, K.; Wang, Y.; Liu, Y. L.; Sankovsky, A.; de Oliveira, C. A.; Mitchell, D. A.; Nizet, V.; McCammon, J. A.; Oldfield, E. Antibacterial Drug Leads Targeting Isoprenoid Biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 123–128.
- (23) Ballester, P. J.; Mangold, M.; Howard, N. I.; Robinson, R. L. M.; Abell, C.; Blumberger, J.; Mitchell, J. B. O. Hierarchical Virtual Screening for the Discovery of New Molecular Scaffolds in Antibacterial Hit Identification. *J. R. Soc. Interface* **2012**, *9*, 3196–3207.
- (24) Barrett-Connor, E.; Mosca, L.; Collins, P.; Geiger, M. J.; Grady, D.; Kornitzer, M.; McNabb, M. A.; Wenger, N. K. Effects of Raloxifene on Cardiovascular Events and Breast Cancer in Postmenopausal Women. *N. Engl. J. Med.* **2006**, *355*, 125–137.
- (25) Jordan, V. C. Fourteenth Gaddum Memorial Lecture - University of Cambridge - January 1993 - a Current View of

- Tamoxifen for the Treatment and Prevention of Breast Cancer. *Br. J. Pharmacol.* **2000**, *131*, 221–231.
- (26) Pfeifer, S.; Fritz, M.; Lobo, R.; McClure, R. D.; Goldberg, J.; Thomas, M.; Pisarska, M.; Widra, E.; Schattman, G.; Licht, M.; Sandlow, J.; Collins, J.; Cedars, M.; Rosen, M.; Vernon, M.; Racowsky, C.; Davis, O.; Dumesic, D.; Odem, R.; Barnhart, K.; Gracia, C.; Catherino, W.; Rebar, R.; La Barbera, A.; Med, A. S. R. Use of Clomiphene Citrate in Infertile Women: A Committee Opinion. *Fertil. Steril.* **2013**, *100*, 341–348.
- (27) Unkila, M.; Kari, S.; Yarkin, E.; Lamminmanta, R. Vaginal Effects of Ospemifene in the Ovariectomized Rat Preclinical Model of Menopause. *J. Steroid Biochem. Mol. Biol.* **2013**, *138*, 107–115.
- (28) Shabsigh, A.; Kang, Y.; Shabsigh, R.; Gonzalez, M.; Liberson, G.; Fisch, H.; Goluboff, E. Clomiphene Citrate Effects on Testosterone/Estrone Ratio in Male Hypogonadism. *J. Sex. Med.* **2005**, *2*, 716–721.
- (29) Durrant, J. D.; McCammon, J. A. Binara: A Novel Algorithm for Ligand-Binding Characterization. *J. Mol. Graphics Modell.* **2011**, *29*, 888–893.
- (30) Trott, O.; Olson, A. J. Autodock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31*, 455–461.
- (31) Manas, E. S.; Xu, Z. B.; Unwalla, R. J.; Somers, W. S. Understanding the Selectivity of Genistein for Human Estrogen Receptor-Beta Using X-Ray Crystallography and Computational Methods. *Structure* **2004**, *12*, 2197–2207.
- (32) Dahlman-Wright, K.; Cavaillès, V.; Fuqua, S. A.; Jordan, V. C.; Katzenellenbogen, J. A.; Korach, K. S.; Maggi, A.; Muramatsu, M.; Parker, M. G.; Gustafsson, J. A. International Union of Pharmacology. Lxiv. Estrogen Receptors. *Pharmacol. Rev.* **2006**, *58*, 773–781.
- (33) Paterni, I.; Bertini, S.; Granchi, C.; Macchia, M.; Minutolo, F. Estrogen Receptor Ligands: A Patent Review Update. *Expert Opin. Ther. Pat.* **2013**, *23*, 1247–1271.
- (34) Minutolo, F.; Macchia, M.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Estrogen Receptor Beta Ligands: Recent Advances and Biomedical Applications. *Med. Res. Rev.* **2011**, *31*, 364–442.
- (35) Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Biomedicine - Defining the "S" in Serms. *Science* **2002**, *295*, 2380–2381.
- (36) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
- (37) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (38) Gee, A. C.; Katzenellenbogen, J. A. Probing Conformational Changes in the Estrogen Receptor: Evidence for a Partially Unfolded Intermediate Facilitating Ligand Binding and Release. *Mol. Endocrinol.* **2001**, *15*, 421–428.
- (39) Katzenellenbogen, J. A. The 2010 Philip S. Portoghesi Medicinal Chemistry Lectureship: Addressing the "Core Issue" in the Design of Estrogen Receptor Ligands. *J. Med. Chem.* **2011**, *54*, 5271–5282.
- (40) Nettles, K. W.; Bruning, J. B.; Gil, G.; O'Neill, E. E.; Nowak, J.; Hughs, A.; Kim, Y.; DeSombre, E. R.; Dillis, R.; Hanson, R. N.; Joachimiak, A.; Greene, G. L. Structural Plasticity in the Oestrogen Receptor Ligand-Binding Domain. *EMBO Rep.* **2007**, *8*, 563–568.
- (41) Durrant, J. D.; McCammon, J. A. Computer-Aided Drug-Discovery Techniques That Account for Receptor Flexibility. *Curr. Opin. Pharmacol.* **2010**, *10*, 770–774.
- (42) Carlson, K. E.; Choi, I.; Gee, A.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Altered Ligand Binding Properties and Enhanced Stability of a Constitutively Active Estrogen Receptor: Evidence That an Open Pocket Conformation Is Required for Ligand Interaction. *Biochemistry* **1997**, *36*, 14897–14905.
- (43) Chen, Z.; Tian, G. H.; Wang, Z.; Jiang, H. L.; Shen, J. S.; Zhu, W. L. Multiple Pharmacophore Models Combined with Molecular Docking: A Reliable Way for Efficiently Identifying Novel Pde4 Inhibitors with High Structural Diversity. *J. Chem. Inf. Model.* **2010**, *50*, 615–625.
- (44) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (45) Langdon, S. R.; Blagg, J.; Brown, N. Scaffold Diversity in Medicinal Chemistry Space. In *Scaffold Hopping in Medicinal Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA, 2013; pp 39–60.
- (46) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742–1753.
- (47) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, *51*, 2174–2185.
- (48) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs 0.1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (49) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.
- (50) Brzozowski, A. M.; Pike, A. C. W.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular Basis of Agonism and Antagonism in the Oestrogen Receptor. *Nature* **1997**, *389*, 753–758.
- (51) Durrant, J. D.; Amaro, R. E. Machine-Learning Techniques Applied to Antibacterial Drug Discovery. *Chem. Biol. Drug Des.* **2015**, *85*, 14–21.
- (52) Katzenellenbogen, J. A.; Johnson, H. J.; Myers, H. N. Photoaffinity Labels for Estrogen Binding Proteins of Rat Uterus. *Biochemistry* **1973**, *12*, 4085–4092.
- (53) De Angelis, M.; Stossi, F.; Carlson, K. A.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Indazole Estrogens: Highly Selective Ligands for the Estrogen Receptor Beta. *J. Med. Chem.* **2005**, *48*, 1132–1144.
- (54) Rdkit: Open-Source Cheminformatics. <http://www.rdkit.org/>.

Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands

Jacob D. Durrant¹, Kathryn E. Carlson², Teresa A. Martin², Tavina L. Offutt¹, Christopher G. Mayne², John A. Katzenellenbogen², and Rommie E. Amaro^{1*}

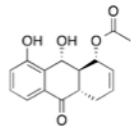
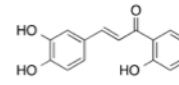
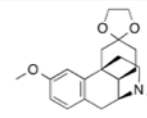
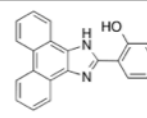
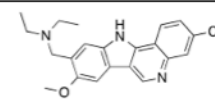
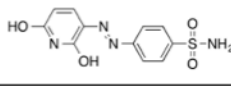
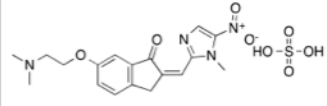
¹Department of Chemistry & Biochemistry and the National Biomedical Computation Resource, University of California, San Diego, La Jolla, CA, 92093. ²Department of Chemistry, University of Illinois at Urbana-Champaign, Champaign, IL, 61801.

* Corresponding author: ramaro@ucsd.edu

|

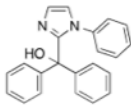
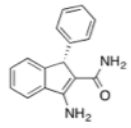
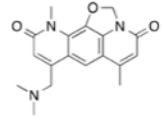
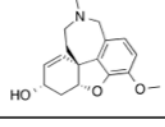
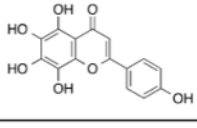
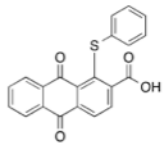
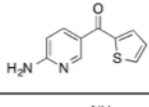
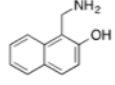
Additional Experimentally Validated ER Inhibitors

Table S1. Additional compounds docked and evaluated using HTVS-SP-XP-NN2 into an ER $_{\alpha}$ structure in the antagonist-bound conformation.¹

Compound	Structure	ER $_{\alpha}$ K $_i$ (μ M)	ER $_{\beta}$ K $_i$ (μ M)	HTVS-SP-XP-NN2
NCI-118628		3.28	3.39	8.50
NCI-37433		3.45	3.45	8.53
NCI-116397		3.45	3.85	8.55
NCI-332670		4.76	3.64	7.87
NCI-317605		6.25	6.06	7.98
NCI-134199		6.67	20.00	8.03
NCI-277184		7.69	5.88	8.08

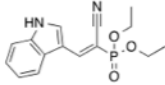
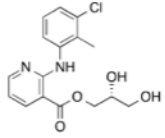
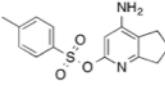
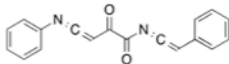
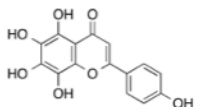
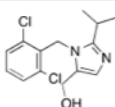
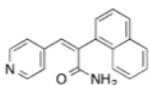
Note that the compounds themselves were tested only for binding, not for agonism vs. antagonism. Docking scores are meant to represent pK $_i$, with higher scores corresponding to more potent binders.

Table S2. Additional compounds docked and evaluated using HTVS into an ER α structure in the agonist-bound conformation.²

Compound	Structure	ER α K $_i$ (μ M)	ER β K $_i$ (μ M)	HTVS
NCI-40269		3.23	3.28	-8.36
NCI-108750		4.17	3.77	-8.90
NCI-275428		5.13	4.76	-8.42
NCI-100058		5.26	4.55	-8.34
NCI-76988		5.56	3.28	-8.94
NCI-117987		5.71	4.35	-8.46
NCI-343344		6.06	3.85	-8.36
NCI-48422		6.45	5.41	-8.54

Note that the compounds themselves were tested only for binding, not for agonism vs. antagonism. Docking scores are in kcal/mol, with lower scores corresponding to more potent binders.

Table S3. Additional compounds docked and evaluated using HTVS-SP-XP-NN1 into an ER α structure in the agonist-bound conformation.²

Compound	Structure	ER α K $_i$ (μ M)	ER β K $_i$ (μ M)	HTVS-SP-XP-NN1
NCI-660151		4.76	7.69	0.88
NCI-335504		5.00	5.41	0.85
NCI-3753		5.26	3.92	0.78
NCI-43308		5.41	5.00	0.93
NCI-76988		5.56	3.28	0.80
NCI-269904		5.56	4.76	0.91
NCI-144694		6.25	5.26	0.95

Note that the compounds themselves were tested only for binding, not for agonism vs. antagonism. Higher docking scores correspond to more potent binders.

Full Experimental Results

Table S4. Relative binding affinity (RBA) values.

Ligand	RBA ER α	RBA ER β	β/α
NCI-3753	0.0038 \pm 0.001	0.0051 \pm 0.001	1.8
NCI-12262	0.0069 \pm 0.003	0.0042 \pm 0	0.61
NCI-13151	0.0148 \pm 0.003	0.0331 \pm 0.006	2.2
NCI-17128	0.0114 \pm 0.002	0.0106 \pm 0.003	0.93
NCI-19136	0.0439 \pm 0.0004	0.0256 \pm 0.005	0.58
NCI-33005	0.026 \pm 0.005	0.154 \pm 0.028	5.9
NCI-34875	0.009 \pm 0.002	0.020 \pm 0.005	2.2
Naringenin	UIUC previous data	UIUC previous data	UIUC previous data
	0.0069	0.0135	2.0
NCI-36586 Genistein	0.024 \pm 0.009	6.14 \pm 1.6	256
	UIUC previous data	UIUC previous data	UIUC previous data
	0.027	6.06	224
NCI-37433	0.0058 \pm 0.002	0.0058 \pm 0.0004	1.0
NCI-40269	0.0062 \pm 0.0003	0.0061 \pm 0.001	0.98
NCI-43308	0.0037 \pm 0.001	0.0040 \pm 0.001	1.1
NCI-48422	0.0031 \pm 0.001	0.0037 \pm 0.001	1.2
NCI-76988	0.0036 \pm 0.001	0.0061 \pm 0.0006	1.7
NCI-78623	0.0076 \pm 0.002	0.0127 \pm 0.002	1.7
NCI-95909	0.0055 \pm 0.001	0.0068 \pm 0.001	1.2
NCI-100058	0.0038 \pm 0.0003	0.0044 \pm 0.0011	1.2
NCI-108750	0.0048 \pm 0.001	0.0053 \pm 0.001	1.1
NCI-111847	0.0037 \pm 0.0003	0.0068 \pm 0.0005	1.8
NCI-112541	0.0050 \pm 0.001	0.0098 \pm 0.002	2.0
NCI-116397	0.0058 \pm 0.002	0.0052 \pm 0.001	0.90
NCI-117544	0.0039 \pm 0.001	0.0068 \pm 0.001	1.7
NCI-117987	0.0035 \pm 0.001	0.0046 \pm 0.001	1.3
NCI-118628	0.0061 \pm 0.001	0.0059 \pm 0.002	0.97
NCI-122253	0.0102 \pm 0.001	0.0096 \pm 0.002	0.94
NCI-130847	0.0098 \pm 0.001	0.0076 \pm 0.002	0.78
NCI-134199	0.0030 \pm 0	0.0010 \pm 0.0001	0.33
NCI-144694	0.0032 \pm 0.001	0.0038 \pm 0.001	1.2
NCI-165701	0.0081 \pm 0	0.0165 \pm 0.005	2.0
NCI-201863	0.0063 \pm 0.0004	0.0072 \pm 0.0004	1.1
NCI-246999	0.0048 \pm 0.001	0.0133 \pm 0.0005	2.8
NCI-269904	0.0036 \pm 0.0003	0.0042 \pm 0.001	1.2
NCI-275428	0.0039 \pm 0.001	0.0042 \pm 0.001	1.1
NCI-277184	0.0026 \pm 0.0001	0.0034 \pm 0.0004	1.3
NCI-308849	0.0146 \pm 0.001	0.0104 \pm 0.0002	0.71
NCI-317605	0.0032 \pm 0.001	0.0033 \pm 0.001	1.0
NCI-319709	0.0040 \pm 0.001	0.0073 \pm 0.001	1.8
NCI-332670	0.0042 \pm 0.0005	0.0055 \pm 0.0002	1.3
NCI-335504	0.0040 \pm 0.001	0.0037 \pm 0.001	0.93
NCI-343344	0.0033 \pm 0.001	0.0052 \pm 0.001	1.6

NCI-351674	0.0073 ± 0.001	0.0062 ± 0.001	0.85
NCI-660151	0.0042 ± 0.001	0.0026 ± 0.0003	0.62

The RBA of estradiol is 100%. Under these conditions, the K_d of estradiol is ~0.2 and ~0.5 nM for ER $_{\alpha}$ and ER $_{\beta}$, respectively. The values shown represent the average plus or minus the standard deviation, calculated using two or more independent measurements. The K_i values reported in Tables 1, 2, and 3 were calculated by dividing the average K_d of estradiol by the RBA, and then multiplying by 100.

Additional Chemoinformatics Analyses

We generated cumulative frequency scaffold plots (CFSPs) to analyze the distribution of compounds over scaffolds.³ To generate these plots, the scaffolds are sorted by their frequency. The cumulative percentage of scaffolds (as a percentage of the total molecules) is then plotted against the cumulative scaffold frequency. An even distribution of compounds appears as a diagonal line and therefore has an area under the curve (AUC) equal to 0.5. If multiple compounds share a common scaffold, the curve is deflected upward. The AUC approaches 1.0 as fewer and fewer scaffolds are required to describe all molecules.

The CFSP AUC values for both the NCI Diversity Set III and the hits identified using each of the three docking protocols were all close to 0.5, regardless of the scaffold method used (Table S5). The scaffolds of the original library were evenly distributed, and the scoring functions did not substantially alter that distribution.

Table S5. Chemical diversity analysis using cumulative frequency scaffold plots.

Compound Database	AUC (Bemis-Murcko)	AUC (Molecular Graph)
NCI Diversity Set III	0.63	0.76
HTVS-SP-XP-NN2/Antagonist	0.53	0.53
HTVS/Agonist	0.53	0.57
HTVS-SP-XP-NN1/Agonist	0.53	0.59

N = total number of compounds in library, AUC = area under the curve from the cumulative frequency scaffold plot.

References

1. Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell*. **1998**, *95*, 927-937.
2. Shiau, A. K.; Barstad, D.; Radek, J. T.; Meyers, M. J.; Nettles, K. W.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A.; Agard, D. A.; Greene, G. L. Structural Characterization of a Subtype-Selective Ligand Reveals a Novel Mode of Estrogen Receptor Antagonism. *Nat. Struct. Biol.* **2002**, *9*, 359-364.
3. Langdon, S. R.; Blagg, J.; Brown, N. Scaffold Diversity in Medicinal Chemistry Space. In *Scaffold Hopping in Medicinal Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA: 2013, pp 39-60.

Chapter 5, in full, is a reprint of the material as it appears in Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands 2015. Durrant, Jacob D.; Carlson, Kathryn E.; Martin, Teresa A.; Offutt, Tavina L.; Mayne, Christopher G.; Katzenellenbogen, John A.; Amaro, Rommie E., J Chem Inf Mod, 2015. The dissertation author was a fourth investigator and author of this paper.