BIOINFORMATIC APPROACHES TO DRUG REPOSITIONING

by

YVONNE YIYUAN LI

B.Sc., The University of British Columbia, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2011

Abstract

Repositioning existing drugs for new therapeutic uses is an efficient approach to drug discovery. However, most successful repositioning cases to date have been serendipitous; the goal of my thesis was to use computational methods to rationally discover drug repositioning candidates.

I first virtually screened (VS) 4621 drugs against 252 drug targets with molecular docking. This method emphasized removing potential false positives using stringent criteria from known interaction docking, consensus scores, and rank information. Published literature indicated experimental evidence for 31 top predicted interactions, supporting the approach. The chemotherapeutic nilotinib was validated as a potent MAPK14 inhibitor *in vitro* (IC_{50} 40nM), suggesting a potential use in inflammatory diseases.

I then applied this method to the cancer target EGFR, predicting the anti-HIV drug tenofovir disoproxil fumarate (TDF) as a novel inhibitor. *In vitro*, TDF inhibited the proliferation and EGFR-signaling of an EGFR-overexpressing cell line, but did not inhibit EGFR in direct kinase binding assays. This study highlighted limitations of computational and experimental methodologies that should be considered when interpreting or designing other studies.

We then screened 1,120 off-patent drugs against the triple-negative breast cancer (TNBC) target p90RSK using both VS and high-throughput (HTS) methods. VS predicted a set of compounds 26-times enriched for known RSK inhibitors and 11 times enriched for HTS hits, underscoring its efficiency. In secondary screens, the chemotherapeutic ellipticine and the bioflavonoids luteolin and apigenin inhibited RSK activity (IC_{50} 0.50-4.77μM), blocked RSK signaling, and inhibited TNBC cell proliferation. These drugs thus have potential to be repositioned to TNBC.

Finally, we rationally repositioned renal cell carcinoma drugs for a patient with a rare tongue adenocarcinoma. Whole genome and transcriptome sequencing of the patient's tumor and normal cells detected sequence, copy number, and expression aberrations, and analysis

suggested that the tumor was driven by the RET oncogene. Treatment with RET-inhibiting drugs stabilized the disease for eight months, after which the disease progressed. We also sequenced the post-treatment tumor and found changes consistent with acquired therapeutic resistance.

Overall, this thesis details two novel high-throughput approaches for drug repositioning: virtual screening of drugs and targets and personalized medicine via sequencing.

Preface

The work presented in this thesis included contributions from many, as mentioned below.

I conceived and conducted the work in Chapter 2, guided by my supervisor, Steven Jones. Jianghong An aided the method design discussions, especially regarding technical issues of the ICM software. I drafted the manuscript, which was edited and refined by Steven Jones and Jianghong An.

The work presented in Chapter 3 was a combination of computational prediction and biological validation. I conducted all the analysis and labwork and Dr. Steven Jones supervised the overall study. Dr. Sandra Dunn at the Child and Family Research Institute guided the experimental design. Anna Stratford aided me in learning and performing the labwork. Kaiji Hu operated the Cellomics high-content screening machinery. I drafted the manuscript; Drs. Jones, Dunn, and Stratford helped edit and refine the text.

Chapter 4 was also in collaboration with Dr. Dunn's lab. I conceived and conducted all the computational analysis in this chapter, supervised by Drs. Jones and Dunn. Jennifer Law, Kristin Reipas, and Amarpal Cheema performed the experimental aspects of the study. I also aided in preparing and revising the manuscript.

Chapter 5 is the result of a large collaboration between Canada's Michael Smith Genome Sciences Centre, led by Steven Jones and Marco Marra, and Janessa Laskin, a clinician scientist at the BC Cancer Agency. They participated in the experimental design, analysis and drafting the manuscript. Ethics approval for this research was granted by the BC Cancer Agency Research Ethids Board, certificate number: H10-01869. I designed and performed all the computational drug analysis under the guidance of my supervisor, conceiving the list of recommended therapeutics (Table 5.3, 5.4 and 5.5). I researched the literature to generate hypothetical cancer signaling pathways for the two tumors (Figure 4.3). I also created Figure 5.1. I worked with various datasets obtained GSC members and assisted Obi Griffith with designing the gene expression analysis. Finally, I, Obi Griffith, Yaron Butterfield, Richard

Corbett and Inanc Birol, undertook analysis and aided in manuscript preparation. Jianghong An, Misha Bilenky, Timothee Cezard, Eric Chuah, Anthony Fejes, Malachi Griffith, Ryan Morin, Sohrab Shah, Nina Thiessen, and Richard Varhol contributed to the computational analysis. John Yee, Michael Mayo, Nataliya Melnyk, Margaret Sutcliffe, Jefferson Terry, Thomas Thomson, and David Huntsman contributed to the clinical assessment of the tumor material. Montgomery Martin, Trevor Pugh, Tesa Severson, Angela Tam, Thomas Zeng, Yongjun Zhao, Richard Moore, Martin Hirst and Robert Holt conducted the molecular biology processing and sequencing of the clinical samples.

Publications arising from work presented in this thesis are listed below.

Li, Y. Y., An, J., and Jones, S. J. (2006). A Large-scale Computational Approach to Drug Repositioning. Genome Inform. 17: 239-247.

Li, Y. Y., An, J., and Jones, S. J. M. (2011). A Large-scale Computational Approach to Finding Novel Targets for Existing Drugs. PloS Comput. Biol 7(9): e1002139.

Stratford, A. L., Fry, C. J., Desilets, C., Davies, A.H., Cho, Y. Y., Li, Y., Dong, Z., Berquin, I.M., Roux, P. P., and Dunn, S. E. (2008). Y-box binding protein-1 serine 102 is a downstream target of p90 ribosomal S6 kinase in basal-like breast cancer cells. Breast Cancer Res. 2008;10(6):R99

Law, J. H., Li, Y., To, K., Wang, M., Astanehe, A., Lambie, K., Dhillon J., Jones, S. J. M., Gleave, M. E., Eaves, C. J., and Dunn, S. E. (2010). Molecular Decoy to the Y-Box Binding Protein-1 Suppresses the Growth of Breast and Prostate Cancer Cells whilst Sparing Normal Cell Viability. PloS ONE 5(9): et12661.

Law, J. H., Reipas, K. M., Cheema, A. S., Li, Y., Li, H., Cherkasov, A., Jones, S., and Dunn, S. E. (2011). Drug repositioning identified p90 ribosomal S6 kinase inhibitors that block triple-negative breast cancer and tumour-initiating cell growth. Submitted.

Jones, S. J., Laskin, J., Li, Y. Y., Griffith, O. L., An, J., Bilenky, M., Butterfield, Y. S., Cezard, T., Chuah, E., Corbett, R., *et al*. (2010). Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. Genome Biol. 11: R82.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADMET — *Absorption, Distribution, Metabolism, Excretion, Toxicity*. An acronym used in pharmacology that describes important properties of a drug when entering the human body.

BIM-8 — *Bisindolyl maleimide-based inhibitor*. One of a series of nanomolar inhibitors of Protein Kinase C that were found to be less specific than previously thought. In Chapter 2, I show the computational method was able to predict the known targets of this drug.

CML — *Chronic Myeloid Leukemia*. A cancer of the white blood cells that has become a paradigm for rational drug design ever since the discovery that 1) the disease is caused by a single translocation event fusing the BCR and ABL genes and 2) the small molecule drug imatinib potently inhibits ABL and is an effective therapy in the clinic.

EF — *Enrichment Factor*. A metric used in virtual screening to assess the ability to predict known interactions. It is calculated as the ratio of the percentage of known ligands among the top X% of docking predictions compared to the percentage of known ligands among the entire database.

EGFR — *Epidermal Growth Factor Receptor*. Also called ERBB1 or HER1. A receptor tyrosine kinase that responds to growth factors and initiates a MAPK signaling cascade. Overexpression of and mutations in EGFR have been observed in many cancers. EGFR is the main focus of Chapter 3, where I aim to find existing drugs that can also inhibit this target.

HMM — *Hidden Markov Model*. A probabilistic method for randomly generating observable data based on an underlying model of discrete states with transition probabilities. An HMM method [341] was used in Chapter 5 to

segment the tumor genomes into regions of loss, neutral, gain, and multiple gain(s) compared to the reference genome.

HTS — *High-Throughput Screening*. A technique where large libraries of chemicals are tested in experimental assays to identify chemicals that have a desired biological activity (i.e. inhibit a drug target or kill cancer cells).

ICM — *Internal Coordinate Mechanics*. The commercial molecular modeling and environment package used throughout this thesis. It is developed by Molsoft, LLC, and has consistently performed well in benchmark studies. In this thesis, we use it to perform multiple sequence alignment, docking, virtual screening, as well as loop and homology modeling.

MAPK14 — *Mitogen-activated proten kinase 14*. Also called p38-alpha. A member of the MAPK family of signaling proteins, this particular enzyme is activated by environmental stress and pro-inflammatory cytokines. It is a drug target under investigation for inflammation-related diseases. In Chapter 2, we found two existing drugs (inflammation-unrelated) that can inhibit MAPK14 activity.

MC — *Monte-Carlo*. A computational algorithm that uses repeated random sampling to model complex systems. These systems would be infeasible if calculated using deterministic algorithms. The algorithm involves generating a random input, and a local deterministic calculation, and performing a set number of iterations. It is the heuristic method used by the molecular docking software ICM, which is described in Chapter 1.

PDB — *Protein Data Bank*. The largest public repository for experimentally determined protein three-dimensional structures, as submitted by researchers worldwide.

RMSD	*Root Mean Square Deviation*. A measure of the difference between values predicted by a model and values actually observed. It is frequently used in docking studies to measure the distance between the ligand conformation predicted by docking and its experimental structure in PDB.
SAR	*Structure-Activity Relationship*. The relationship between the chemical structure properties of a molecule and its biological activity. In drug discovery, derivatives of active chemicals are synthesized and retested against the target, in order to determine chemical groups and positions that are important for binding affinity.
SBDD	*Structure-Based Drug Design*. A strategy in drug discovery where drugs are design based on the three dimensional structure of the protein target. Knowing the important residues in the binding site allows for more rational design of binders.
TDF	*Tenofovir Disoproxil Fumarate*. A drug marketed by Gilead Sciences (Foster City, CA) that is indicated for people with human immunodeficiency virus (HIV). It was predicted as a potential inhibitor of the Epidermal Growth Factor Receptor (EGFR).
TNBC	*Triple Negative Breast Cancer*. A subtype of breast cancer defined as tumors lacking expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER-2). Compared to other breast tumors, TNBCs are more aggressive, with higher rates of recurrence and death. Currently, there are no established targeted therapies for this subtype.
VS	*Virtual Screening*. A computational technique where large libraries of chemicals are rapidly assessed to identify chemicals that are most likely to bind to a drug target.

YB-1 *Y-box binding protein 1*. A transcription factor that is directly activated by RSK and subsequently induces transcriptions of multiple oncogenes, including EGFR and MET. We indirectly targeted YB-1 in Chapter 4 by finding inhibitors of RSK.

Acknowledgements

First and foremost, I would like to give my sincerest thanks to my supervisor, Dr. Steven Jones, for his encouragement, support, and guidance ever since my first foray into bioinformatics as a Co-op student and all throughout my graduate studies. I am thankful for the opportunity to work at the Genome Sciences Centre, as well as to go to stimulating scientific conferences to present my research. Thanks to other members of the drug discovery group at GSC, Dr. Jianghong An and Dr. Alex Yakovenko for interesting project discussions. A special thanks to Dr. An for obtaining and setting up the huge docking cluster as well as helping me with all my docking related questions throughout my graduate studies.

To Dr. Sandra Dunn and the members of the Dunn Lab, I would also like to express my gratitude for welcoming me to their lab. Through our collaboration in the last two and half years, Dr. Dunn has always been supportive and encouraging. A special thanks to Dr. Anna Stratford for teaching me all the basics of cell culture and helping me with even the most basic of experiments. Thanks as well to Jennifer Law for her collaboration and patience throughout the RSK and YB-1 projects. These projects were very interesting computational endeavors, and I thank to Dr. Dunn for the opportunity to work on it.

I am grateful for the support and guidance of my thesis advisory committee, Drs. Artem Cherkasov, Marco Marra, and Marianne Sadar for their expert opinions and advice from both computational and experimental viewpoints, as well as for their efforts reviewing this thesis. To the administrative staff, Louise Clarke, Lulu Crisostomo, Jillian McKenna, and Sharon Ruschkwoski, I am also thankful for all their help and patience throughout these years.

I extend my thanks to all members of the BC Cancer Agency's Genome Sciences Centre, including the staff, scientists, and fellow graduate students for creating such a thriving scientific environment. In particular I thank fellow graduate students past and present for insightful and entertaining discussions, including Anthony Fejes, Olena Morozova, Elizabeth Chun, Sorana Morissey, Obi Griffith, Monica Sleumer, Jacqueline Lai, and Simon Chan.

I would like to acknowledge all the funding sources throughout my graduate studies including the Natural Sciences and Engineering Research Council of Canada, the Bioinformatics Training Program, and my supervisor Dr. Jones.

Finally, to my parents who have supported me wholly every day of my life, I would like to express my deepest gratitude for raising me well and taking such good care of me. I also thank my large extended family for their steady care and support from all over the world.

1 Introduction

1.1 History of drug discovery and motivation for drug repositioning

Drug discovery is the process by which chemical substances are developed to treat diseases. Historically, natural product extracts from plants, animals, and minerals were used. With progress in chemistry in the 19th century, researchers started isolating active substances from the extracts: the precursor of aspirin was extracted from willow tree bark [3], the antibiotic penicillin from bread mold [4], and the painkillers morphine and codeine from the opium poppy [5]. However, finding effective natural sources relied heavily upon serendipity. For instance, the discovery of dicoumarol as an anticoagulant stemmed from farmers finding that their cattle died from internal hemorrhaging after ingesting rotten sweet clover [6]. The advent of biochemistry in the 1930's introduced the concepts of enzymes and receptors and their potential as drug targets, leading to targeted drugs like monoclonal antibodies [7]. The field of molecular biology brought greater understanding of disease mechanisms at the molecular level, such as the cancer causing mutations in RAS and P53, overexpression of HER2, and translocation of BCR and ABL [8]. This knowledge has led to a more rational approach towards drug discovery, where drug candidates are obtained by screening chemicals against a target known to play a role in disease. The development of X-ray crystallography has allowed us to determine the 3-dimensional structures of proteins at atomic resolution, launching the more recent field of structure-based-drug-design (SBDD). Notable drugs that have been developed through SBDD approaches include Viracept for AIDS and Relenza for influenza [9]. With the completion of the Human Genome Project leading to an increase in the number of drug targets, modern drug discovery has also expanded to include quality-of-life drugs including sildenafil for erectile dysfunction and statins for obesity.

The rate of new drug approval has remained relatively constant in the past 60 years, with just 20-30 new drugs approved per year [10]. However, recent productivity charts have shown tremendous increases in expenditure. In 2006, large pharmaceutical companies spent $92 billion researching and developing only 22 new drugs [11]. The average time and cost to discover one successful new drug and bring it to market is currently estimated at $1.78

billion USD and 13.5 years [12] – a staggering value that does not even include the time and cost involved to identify and validate the target. In the standard drug discovery pipeline (Figure 1.1a), clinical trials development is the most time consuming (5-6 years) and expensive step (63% of the overall cost [12]). Moreover, only 11% of potential drugs entering clinical trials have a chance of successfully continuing on to US Food & Drug Administration (FDA) approval [13]. The inefficiency of pharmaceutical drug development has been widely discussed [10-18].

Drug repositioning is the process of finding new therapeutic uses for existing drugs, outside of their original medical indications. This is an efficient approach to discovery (Figure 1.1b), since existing drugs are already optimized to their target, have extensive absorption, distribution, metabolism, excretion, and toxicity (ADMET) data, and are less likely to fail clinical trials due to adverse effects [15]. To date, most repositioned drugs have been discovered through serendipitous observations: sildenafil was first developed for angina but later approved for erectile dysfunction; thalidomide was first marketed for morning sickness and later for leprosy and multiple myeloma [15]. Rationally repositioned drugs also exist. The most well known example is imatinib, first approved for chronic myeloid leukemia (CML) due to its inhibition of the BCR-ABL fusion protein and later approved for gastrointestinal stromal tumor due to its potent inhibition of KIT [19]. A second example is duloxetine, which was developed to treat depression but later marketed for stress urinary incontinence based on a shared mechanism of action between the two diseases [15]. With the increasing catalog of validated drug targets and drug molecules, opportunities for elucidating novel drug-target, drug-mechanism, or target-disease relationships are also rising. These relationships form a basis from which repositioning hypotheses can be made.

1.2 Drug discovery

1.2.1 Drugs and drug targets

For purposes of this thesis, drugs are molecular structures that can alter the biological activity of living systems for medicinal purposes. Targets are molecular structures that interact with these drugs, and their change in activity upon drug-binding is used for medicinal purposes

[20]. These medicinal purposes encompass the treating, diagnosing, and preventing of disease.

There are many types of drugs currently being developed: small molecules, peptides, proteins, antibodies, oligonucleotides, and aptamers. Drugs approved by the FDA since the 1950's have predominantly been small molecules, accounting for 1,103 of 1,222 drugs [10]. Antibodies comprise the majority of the other 119, but while they are highly specific to their target, they only target cell surface and secreted proteins. In addition, the large size of antibodies (approximately 150kDa) compared to small molecule drugs (approximately 500 Da) introduces difficulties in tissue penetration, blood clearance, and high production costs [21, 22]. Of small molecule drugs, inhibitors are a much more attractive option than agonists because designing a chemical to cause loss-of-function is a much simpler task than causing gain-of-function. Small molecule drugs also tend to target protein active sites or cofactor-binding sites rather than protein-protein binding sites, as the latter are generally flat and open [23] and form binding interactions through large interfacial areas with many hydrophobic residues [24]. In contrast, protein-ligand binding usually involves a deep cavity as well as a few strong hydrogen bonds and electrostatic interactions.

Four types of drug targets are targetable by small molecule drugs: proteins, nucleic acids, polysaccharides, and lipids. Much effort has been invested in studying the 'druggable genome' – the proteins in the genome that have the potential to interact with drugs. A ligand-binding analysis in 2002 estimated 399 existing targets and ~3000 druggable targets [25]. In 2007, Imming *et al*. estimated 218 drug targets based on marketed drug information [20]. The latest version of DrugBank, a curated open-access database with drug-target interactions culled from the literature, contains 4,326 unique targets, of which 1,768 are targets of FDA approved drugs. However, it should be noted that the DrugBank targets are not necessarily involved in disease, but just physically interact with the drugs. Within the established drug targets, protein families with more than 40% of marketed drugs include G-protein coupled receptors and ion channels [26]. Kinases are a popular target for cancer therapy – kinase inhibitors being the largest class of new cancer drugs – not only due to their involvement in cell growth, proliferation, and survival, but also because they are frequently mutated in

cancers [27]. Establishing a protein as a therapeutically relevant drug target is a difficult task due to the tremendous amount of *in vitro* and i*n vivo* research involved. There is added challenge in proving the effectiveness of a target in terms of the safety and efficacy profiles in clinical trials. On average, only 5.3 new targets accompany approved drugs each year [26].

1.2.2 Fundamental approaches to drug discovery

Classical drug discovery employs a forward pharmacology approach. Compounds with biological activity are first discovered, and extensive research follows to determine the compounds' molecular targets and mechanism of action [28]. For example, siacylic acid (the precursor of aspirin) was used to treat arthritis in the 1870's, but its inhibition of cyclooxygenases was not revealed until the late 1970's [29]. Another instance is the class of fluoroquinolone antimicrobial agents. The first generation compound nalidixic acid was discovered in 1962 and its DNA gyrase inhibition mechanism was delineated ten years later [30]. It was not until the 1990's that the other major target of these compounds was determined to be topoisomerase IV [31]. One current approach to forward pharmacology approach is high-content screening (HCS), where phenotypic screening of compound libraries or natural product extracts is conducted on cell lines or other organisms with quantifiable phenotype changes [32]. Subsequent determination of how a drug works at the molecular level is a challenging task; to date, there are at least 30 marketed drugs with unknown mechanisms of action [20].

The reverse pharmacology approach relies on first identifying a molecular target, and then finding a drug that affects the target's molecular function [28]. An example of this approach is high-throughput screening (HTS), where large chemical repositories are tested against a specific target *in vitro*, to find chemicals that modulate the target's biological activity. The majority of HTS systems assay for a purified protein's enzymatic activity, though HTS can also measure the activity of signaling transduction pathways either isolated or in cellular environments. Current technology can test 100,000 compounds per day using automated robotic systems [33]. Compounds showing activity then undergo validation assays and animal model studies, to confirm whether their biological activities are indeed due to inhibition of the target. Reverse discovery has been the conventional strategy for the past few

decades; however, this approach requires a deep understanding of the biological activity of target proteins.

1.2.3 From the magic bullet to the multi-target paradigm

The concept of 'magic bullets,' drugs that bind directly to a single molecular disease target, was first postulated by Paul Erlich in the late 19th century [34]. This "one-drug one-target one-disease" strategy has driven much of drug discovery in the late 20th century, and has resulted in successful targeted therapies. The most well known examples are the antibody Herceptin for HER2-positive breast cancer, the antibody Rituxan for non-Hodgkin's lymphoma, and the small molecule imatinib for CML [35]. For diseases where a single protein is known to be the unique driving aberration - such as the fusion protein BCR-ABL in CML [36] - the monotherapy approach has proven effective. However, many diseases are caused by multiple molecular aberrations, including certain cancers, noninsulin dependent diabetes mellitus, and Alzheimer's [37]. In addition, the majority of human cancers exhibit extensive molecular and phenotypic heterogeneity [38]. For these 'complex' diseases, where the cause is not any single defect, multiple proteins must be targeted concurrently.

There are two types of multi-targeting approaches. The first is to use a combination therapy of multiple approved drugs. This strategy has already been implemented for acquired immunodeficiency syndrome (AIDS) through HAART (highly active antiretroviral therapy). One widely used treatment, ATRIPLA, is a cocktail of the nucleotide reverse transcriptase inhibitor (RTI) tenofovir disoproxil fumarate, the nucleoside RTI emtricitabine, and the non-nucleoside RTI efavirenz [39] – three inhibitors that target the human immunodeficiency virus (HIV) through distinct mechanisms. The second multi-targeting approach is to use a single drug that simultaneously inhibits several different targets – that is, a drug with clinically relevant polypharmacology [26]. Many antipsychotic drugs fall into this category, such as Clozaril, a multi-targeting drug used to treat schizophrenia. Unexpectedly, Clozaril showed less efficacy when chemical modifications were made to improve its specificity [40]. Sunitinib is a kinase cancer drug with a large number of targets known to be involved in cell proliferation, angiogenesis, and the tumor microenvironment. It was found to be more effective than single-target drugs in mouse xenograft models, and had a cumulative anti-

tumor efficacy similar to combining the single-target drugs [41]. These examples have demonstrated that targeting multiple proteins is an effective strategy for complex diseases.

1.2.4 Drug polypharmacology

A growing body of evidence suggests that small molecule drugs have extensive polypharmacologies. This is well illustrated in the case of ATP-competitive inhibitors of the kinase family, since every kinase has an ATP-binding domain. Fabian *et al*. tested 20 kinase inhibitors in ATP-competitive assays against a panel of 119 kinase proteins, and showed that the kinase drugs inhibited many more proteins than expected [42]. Sunitinib at 10μM showed inhibition to 79 of 119 kinases tested, though it had ten-fold stronger binding to its four intended targets than to any other off-target. Vandetanib inhibited 50 of 119 kinases, but showed only two-fold stronger binding to its two intended targets. Brehmer *et al*. washed cell lysate extracts over a bead column fixed with gefitinib, and discovered 26 proteins binding to this EGFR-specific drug [43]. The non-steroidal anti-inflammatory drug (NSAID) celecoxib is a selective COX-2 inhibitor, preventing the production of prostaglandins that cause pain and inflammation. It also is a nanomolar inhibitor of carbonic anhydrase II [44] and exhibits *in vitro* and *in vivo* inhibition of 5-lipoxygenase (5-LO), though at a micromolar potency [45]. This is may account for some of the COX-2-independent effects of celecoxib, as 5-LO is also involved in inflammation through a parallel pathway to COX-2 [46].

Drug polypharmacology can also be an undesirable phenomenon. A number of diverse drugs, including the serotonergic 5-HT4 receptor agonist cisapride, the histamine H1 receptor inhibitors astemizole and terfenadine, and the antibacterial drug grepafloxacin, were withdrawn from the market due to causing an increased risk of life-threatening ventricular arrhythmias [47]. This adverse effect was found to be due inhibition of hERG potassium channel, a key protein in cardiac repolarization and a target shared between all four drugs.

In light of this evidence, several recent studies have analyzed the drug-target space. Paolini *et al*. created a human pharmacology interaction network connecting proteins that have one or more chemical binders in common (Figure 1.2) [48]. In their database of 276,122 active compounds, 35% were observed to hit more than one target. Though the majority of

compounds only bound to targets in the same gene family, 25% were 'promiscuous' compounds that also bound targets from different gene families. A second study by Yildirim *et al*. mapped the network for existing drugs and found 305 out of the 395 drug targets were linked by multi-targeting drugs [49]. Mestres *et al*. consolidated seven drug-target interaction databases, creating an interaction network between 802 drugs and 480 targets, and found that on average each drug interacted with six different targets [50].

During the drug development process, candidate drugs are routinely tested against a small panel of proteins that are similar to their intended target. However, more and more studies are demonstrating that small protein panels are no longer sufficient for assessing drug specificity, particularly in the case of kinase drugs. Though the prospect of finding multi-targeting drugs is attractive, the actual implementation is a more complicated endeavour. Drugs will have to be screened against multiple targets at a time, and specific combinations of target affinities must be attained. Though high-throughput assays for over 317 kinases have been developed [51], assays are lacking for other protein families. An even more challenging task is determining which set of targets will be optimal in a given disease. In short, to approach multi-target drug discovery in a rational manner, more information about targets and their pathways in diseases must first be elucidated.

1.3 Computational drug discovery

Computational methods are an integral part of drug discovery. Broadly, they are used to predict novel drug targets, protein-ligand binding, perform virtual library screening to find novel compounds, optimize the efficacy of lead compounds, perform *de novo* design of novel drugs, predict drug properties like ADMET or potential drug-drug interactions [52]. Even for experimental methods like X-ray crystallography or high-throughput assays, computational methods are necessary for structure refinement or data collection and correction. Much of the research conducted in the field of computer-based drug discovery (CBDD) overlaps with chemoinformatics, the field where informatics methods are applied to chemistry problems. In this section, computational methods for finding novel drug candidates are discussed, with particular focus on molecular docking methods.

1.3.1 Overview of current approaches

There are two main types of computational methods for finding novel ligands for protein targets: ligand-based and structure-based methods. Ligand-based methods rely on the knowledge of a set of ligands that bind to a biological target of interest. From that list, a quantitative structure-activity relationship (QSAR) can be derived between activity and ligand structure (2D, 3D, or ligand molecular properties such as charge or melting point) [53]. Pharmacophore models can also be constructed relating ligand activity to their functional groups [52]. If the 3-dimensional structure of a protein is available, structure-based methods can be applied. One technique is molecular docking, where a small molecule ligand is rendered flexible and fitted into a protein binding site in 3-dimensional space. The 'best fitting' conformation of the ligand is determined, and the likelihood of the binding interaction is reported as a score. Beyond docking, molecular dynamics is a computationally intensive method that uses classical force fields to model a protein-ligand binding in full aqueous surroundings, sampling of all degrees of freedom, for a snapshot in time (nanoseconds) [52]. Finally, *de novo* drug design attempts to incrementally construct a chemical that can inhibit the target binding site from scratch [54]. This can result in novel chemotypes and binding scaffolds; however, the predicted chemicals are not always readily synthesizable. Of the three methods, docking is the most common method used in high-throughput virtual screening (VS) due to its high speed, low cost, and software availability. Docking was also the method of choice in this thesis, and is described in detail in the following sections.

1.3.1.1 Virtual screening versus experimental screening

Screening large databases of small molecules to find one that binds to a drug target is analogous to finding a needle in a haystack. Current technologies include experimental (HTS) and computational (VS) approaches, each with benefits and limitations.

Due its computational and predictive nature, VS methods are often regarded as inferior to HTS in the search for novel compounds. However, HTS methods also produce noisy readouts with false positives and false negatives and with significant assay variability.

Sources of HTS errors fall into three categories: logistic, measurement, or strategic errors [55]. Logistic errors are caused by 1) human error such as assigning incorrect information to a compound, 2) intrinsic compound properties such as stability and solubility in solvent, 3) aggregation, 4) purity, or 5) solvent properties such as evaporation during the experiment [55]. In particular, aggregation occurs when compounds form colloidal aggregates during the assay and indiscriminately inhibit enzymes [56]. Also, fluorescent molecules can interfere with the fluorescence reading of the assay. Measurement errors can include imperfect pipetting, temperature gradients, time sequence of processing plates – errors that can be multiplied in high-throughput assays [55, 57]. Lastly, strategic errors can be caused by the assay design, such as whether assay measurements are performed sequentially or in parallel, whether compounds are tested one by one or in a mixture, and inter- or intra- assay variability. Of the artifacts listed, one study found that aggregates caused over 90% of the false positive hits selected by high-throughput assays [58, 59].

One advantage of HTS is the ability to screen natural product extracts, as the compound structures in the extracts are often undetermined and thus not available for docking. Natural products are also attractive to pharmaceutical companies as a source of novel chemical scaffolds and thus potential intellectual property (IP). However, these benefits are balanced by many difficulties leading to high time and cost requirements: 1) natural product compounds often have complex structures that are infeasible to synthesize, 2) natural product sources are limited and may become extinct, 3) the active compound(s) in an extract need to be singled out and elucidated, 4) interesting compounds in the extract may be in concentrations too low to be detected in the screen, and 5) the novel compound discovered may not be patentable due to the IPs of local governments [60, 61]. In comparison, VS methods work with known compound structures for which synthesis methods, toxic chemical moieties, and other drug-like characteristics can be readily assessed.

There are numerous studies comparing virtual and experimental screening. Edwards *et al.* docked ~480,000 small molecules to a GPCR, and then tested a subset of ~4,300 in a high-throughput flow-cytometry platform to measure protein binding [62]. They found a 1.2% hit rate for virtual screening which was 12-fold better than physical screening. Polgar *et al.*

docked ~5300 compounds to GSK3B using FlexX and screened ~16,000 molecules using a robotic AssayStation and found that the VS hit rate of 12.9% was much higher than the 0.55% hit rate of HTS [63]. Paiva *et al*. used the Merck chemical collection for both screenings and found that VS and HTS gave hit rates of 6% and 0.2%, respectively [64]. Doman *et al*. docked 235,000 commercially available compounds to PTP1B and concurrently screened a 400,000 compound corporate library using HTS compounds [65]. The top 365 docked compounds were tested in enzymatic assays and 127 (35%) exhibited some inhibition of PTP1B. In comparison, the HTS screen had a much lower hit rate with only 85 (0.021%) compounds showing inhibition. The two lists of experimentally validated PTP1B inhibitors were very different from each other, suggesting that VS and HTS may be complementary methods. In short, VS is a fast and cost-effective method of screening chemicals, and is a complementary method to HTS. VS has proven to be valuable in selecting a smaller subset of the chemical library with a higher percentage of active compounds. In the following sections, the basics of VS by molecular docking are described in detail.

1.3.2 Molecular docking

Molecular docking is a process that simulates how a small molecule (ligand) interacts with a protein-target binding site (receptor), in terms of a binding conformation and a binding score [1]. A basic docking protocol is shown in Figure 1.3. First, a 3-dimensional structure of the target protein is obtained. Then, the binding site of interest in the protein is designated. Third, a ligand is fitted into the binding site and the most likely ligand binding conformation is selected. Fourth, an overall score is calculated for the protein-ligand docking based on the best conformation. In a virtual screening (VS) procedure, a large compound database is docked one-by-one to the target protein. The scores of all ligands are collected and ranked, after which the most likely binding candidates are selected by visual inspection of the predicted binding conformations. Candidates are then tested experimentally to determine their ability to inhibit the target protein. Molecular docking has been used to successfully discover novel ligands for a wide variety of drug targets (compounds for over 20 targets just from 2007 to mid-2009 are reviewed in [66]).

1.3.2.1 Obtaining a protein 3D structure

The starting point of structure-based drug design is a 3-dimensional protein structure, which can be attained using experimental or computational methods. The two prevalent experimental methods today are X-ray crystallography and nuclear magnetic resonance (NMR), which account for 87% and 12% of existing structures, respectively [67]. In X-ray crystallography, the protein is suspended in solution and crystallized. X-rays are then beamed at the crystals and the scattering pattern is used to determine the electron densities in the crystal. These data are assembled into a 3-dimensional structure, to a resolution of 1.5-3Å [68]. In NMR, strong oscillating magnetic fields are applied to stimulate proton nuclei. The chemical shifts of ^{1}H, ^{13}C, and ^{15}N protons are mapped out and used to reverse engineer the protein structure. These two methods have distinct advantages and disadvantages. X-ray methods can determine the structure of any size of protein to a high resolution as long as suitable crystals can be formed, whereas NMR methods are limited to proteins 5-30 kDa in size and resolutions of 2-10Å but can observe the flexibility and motion of the protein in solution [69]. Though NMR structure resolutions can be improved with computational strategies [70], the general consensus is that X-ray structures are better in quality due to higher experimental data-content and atomic resolution [71].

If the crystal structure of a protein is not available, a comparative (or homology) model of the structure can be built using the existing structure of a similar protein. If the sequence identity between the two proteins is over 50%, an accurate model can be built (~1Å alpha-carbon atom root-mean-square-deviation (RMSD) from an existing experimental structure) [72]. A sequence identity of 30-50% may result in a model with 2-3Å RMSD, while any sequence identity under 30% is not recommended for modeling [72]. Previous analyses have suggested that homology models are useful for virtual screening studies [73, 74]. Specifically, a systematic study showed that docking to homology models is often just as successful as docking to the structure template [75].

A comprehensive resource for protein 3-dimensional structures is the Protein Data Bank [76], which currently houses over 67,000 experimentally determined protein structures [67]. The

number of deposited structures has grown rapidly compared to only 772 structures solved from 1976 to 1992. As more drug target structures become available, the utility of docking in SBDD increases correspondingly.

1.3.2.2 Docking preparation steps

Prior to docking, the quality of the protein structure should be considered. Whether an X-ray structure, NMR structure, or homology model, the resolution of the structure and presence of gaps in the structure are important considerations. Errors in the structure may have occurred during structure deposition and should be noted. Hydrogen atoms must be added to the structure and optimized, since the resolutions of current protein structures are not yet powerful enough to accurately place hydrogen atoms.

The binding sites in a protein can be determined using the knowledge of an existing ligand-bound structure, experimental active site mutation data, evolutionary inference based on sequence alignments of homologous proteins, or a pocket-prediction algorithm on the protein structure. The latter is necessary when a protein structure is not bound to any ligands (a *holo* structure); many algorithms have been developed to address this issue (Q-SiteFinder [77], PocketFinder [78], PASS [79], to name a few). In addition, pocket search can also identify potential allosteric sites (sites other than the active site), which are also viable targets for drugs. For example, imatinib binds to an allosteric site in ABL beside the ATP-binding site to lock the protein in an inactive conformation [80]. Another ABL inhibitor, GNF-2, binds to the myristoyl binding site distant from the ATP-binding site [81, 82].

Compound databases containing millions of compounds are frequently used in VS analyses; they may need to be preprocessed to expedite the docking step during VS. Some docking methods prefer input ligands to be in an energetically minimized conformation [83] or have pre-assigned atom types and charges [84], requiring other software. Also, prior knowledge about binders can be used to pre-filter ligands through similarity to known binders or through a pharmacophore model. Another common filter is Lipinski's rule of 5, a set of descriptors that describe the range of molecular weights, hydrogen bond donors and acceptors, and octanol-water partition coefficients that encompass >90% of existing oral marketed drugs

[85]. Ligands that do not have these properties are more likely to have poor absorption and/or permeation (i.e. do not possess the ideal properties as a drug), and can thus be eliminated from consideration. However, it is important to note that 10% of approved drugs are overlooked with this filter. Thus, Lipinki's rules can be relaxed during VS if computational capacity permits.

With a good quality protein structure, a binding site of interest, and compound structures, docking is ready to begin.

1.3.2.3 Docking and scoring methods [1, 2]

In order to perform VS of millions of compounds, a key criterion for docking processes is speed – on the order of seconds to minutes per simulation. One constraint commonly used in docking is to keep the protein receptor rigid, under the assumption that ligands are often highly flexible and can adapt to complement the receptor site [86]. However, even the potential ligand conformational space can be infinite, depending on the number of rotatable bonds and how many degrees they are rotated at a time. Thus, heuristic methods were developed to provide a fast yet effective sampling of ligand conformational space.

Sampling algorithms fall into two main categories: systematic search algorithms like incremental reconstruction and stochastic methods like genetic algorithms or Monte Carlo (MC) simulations. Incremental reconstruction is the process of separating the ligand molecule into fragments, docking them independently, and merging the fragments back together. This can be done by docking the rigid fragment first then adding on flexible fragments one by one, or by docking each fragment separately and merging them in the binding site. Genetic algorithms generate a collection of starting ligand poses, to which 'mutations' (bond rotation) or 'crossovers' (merging two poses) are applied to form new poses; this process cycles until the poses have converged. The basis of this method is to mimic Darwinian evolution and pick the 'fittest' ligand poses. MC methods start from one single ligand pose, to which random moves like bond rotation and ligand translocation are then applied. The Metropolis criterion is used to evaluate each new pose, and decide whether it is acceptable as the starting pose for the next cycle of modification.

Scoring functions are used to evaluate the docked ligand conformations and fall into three main categories: knowledge-based, empirical, and force-field based. Knowledge-based scoring compares the interactions between the protein and the docked ligand to the known atom-atom interactions across experimental complex structures in PDB. Empirical methods calculate the score as a weighted sum of various terms of protein-ligand interaction potentials such as the hydrogen-bonding potential or the electrostatic potential. The weights for these functions are determined by training on an existing set of PDB structures. Lastly, force-field based methods try to accurately estimate the binding free energy of the protein-ligand interaction, using physics-based energy functions.

1.3.2.4 The ICM docking and scoring method

Here I describe the algorithm of the Internal Coordinate Mechanics (ICM) software [87-90], which was used throughout this thesis to perform docking. Briefly, ICM performs global optimization of an empirical energy function using a Monte Carlo minimization procedure.

A unique feature of ICM is that ligands are represented by an 'internal coordinate system': a start point in the ligand is defined arbitrarily, and 4 types of variables (bond lengths, bond angles, torsion angles, and phase dihedral angles) for each connecting atom of the ligand are defined. Bond lengths and certain bond angles are fixed. In contrast, a Cartesian coordinate system requires 7-10 times more variables including degrees of freedom for bond lengths and angles, atom coordinates, as well as torsion angles, corresponding exponential increase of the ligand conformational space.

At the start of a docking cycle, a random change is applied to the ligand conformation and position (Figure 1.4). This can be a single torsion angle change, a Brownian-like random change in the ligand rotational and translational movement, or a biased probability 'random' torsion change that chooses new torsion angles based on the distribution of existing ligand conformations in PDB. The new ligand conformation is minimized with a simple scoring function, after which a more comprehensive score is determined for the cycle. The Metropolis criterion is applied to accept or reject the conformation, and then the system

restarts with a random conformation change to the ligand. Briefly, the Metropolis criterion states that a new ligand pose with a lower energy should be accepted, but a higher energy pose can be accepted if a probability value (dependent upon the energy difference) is larger than some random number. The leniency of the probability function can be adjusted prior to docking. The cycling stops after a predetermined number of cycles or when the predicted ligand conformations converge at an energetic minimum. Though the MC algorithm tries to avoid falling into local minimum energies by using the Metropolis Criterion, there is no guarantee that the pose is globally minimized. Thus, ICM recommends that dockings should be repeated 2-3 times, and the best scoring pose be retained.

The ICM scoring function is a weighted sum of energy terms. A simpler energy function is used during the local minimization step, with the following ECEPP/3 energy terms:

1. van der Waals energy: two atoms near each other will have binding potential, but will begin to repel each other as they come too close.
2. hydrogen bonding energy: a bonding interaction that forms when a hydrogen atom that is covalently attached to an electronegative atom (such as nitrogen or oxygen) is attracted to another electronegative atom.
3. Coloumb electrostatic energy: an attractive or repulsive interaction that forms when two electrically charged atoms are near each other.
4. torsion energy: the bond energy, bond angle bending, and improper torsion energies in a given ligand conformation.

The more comprehensive scoring function also includes approximations of three energies that are computationally expensive and would be rate-limiting during local minimization.

1. desolvation energy: the energy required to dispel solvent molecules from the binding site and break the protein-solvent interactions in order for ligand to bind.
2. electrostatic polarization energy: the energy caused by reorientation of electronic clouds surrounding protein atoms after ligand binding.
3. side chain entropy: the entropic energy lost by protein side chains when protein atoms are bound to ligand atoms.

When performing VS of many small molecule ligands against a protein receptor, ICM returns a ranked list of the ligands, each with an energy-based score as described above (hereafter referred as the icm-score) and a potential of mean force score (pmf-score). The latter is a knowledge-based score and its statistical-basis indicates that it is entirely independent from the icm-score. ICM recommends the pmf-score for secondary evaluation of the docked ligand, though it is not often utilized in published docking studies.

1.3.2.5 Advantages and limitations of docking & virtual screening methods

Docking methods can help narrow down large chemical libraries to select a smaller subset for experimental testing. Its speed and computational nature allow the screening of compounds much faster and at much lower cost than experimental screening. The 3-dimensional simulation of docking is more thorough than methods that only search for similar binding sites or similar chemical structures. In addition, docking can be used to screen compounds that are not physically available. Finally, docking delivers the predicted conformation of the ligand bound to the protein, which can aid in ligand optimization when combined with SAR analyses

However, docking also has limitations along each step of its process. The first factor is initial protein structure. Though X-ray crystal structures are considered to be 'gold-standard,' differing experimental conditions can result in the ligand orientation to be reversed in the binding site [106]. Certain characteristics of the protein structure may also be due to its crystalline state and may not be representative biologically [68]. Lastly, ambiguities when accounting for structural heterogeneity also contribute to the inaccuracy of crystal structures [107].

The binding pocket itself is a potential caveat. Though docking is often used to find compounds that bind to established ligand-binding sites, sometimes novel allosteric sites need to be predicted and targeted. In such cases, the accuracy of the pocket prediction algorithm is crucial, as results from docking to a non-existent protein pocket are of no utility.

During docking, the protein structure is kept rigid. However, since both the protein and ligand may conform to fit each other during binding, using a protein conformation that is too different from the actual conformation may allow true binders to be missed. Murray *et al.* assessed the effect of receptor flexibility by docking a ligand back to its native crystal structure and also to other structures of the same protein (these other structures were bound to different ligands and thus represented different conformations of the same protein) [108]. They found that the success rate of docking decreased from 76% for the native scenario to 49% in the flexible scenario.

The heuristic sampling method is inherently a limitation since the optimal binding pose can be missed. Furthermore, no water molecules or co-factor molecules are taken into account during docking unless they are fixed into the binding site beforehand. The aqueous environment is important as water molecules often mediate interactions between the protein and ligand.

Limitations of scoring functions are also an important issue in molecular docking. During docking, inaccurate energy functions can cause the selection of the wrong ligand conformation as the best-docked pose. During scoring, empirical functions may fail to score accurately if the ligand does not fit the trained score function, and force-field methods may fail if the force field parameters are not complete. Even state-of-the-art scoring functions cannot correctly predict the binding energies [109]; however, scoring functions are more effective when used to select likely binders for further validation, instead of predicting precise binding affinities.

Finally, though docking may model the binding interaction that occurs in an *in vitro* binding assay (without protein flexibility, or an aqueous environment), it is still much simpler than the binding process when a compound enters a cell. In the latter scenario, the compound may have hundreds of thousands of protein binding sites to choose from, some of which (like ATP-binding sites) may be very similar to each other.

Overall, these weaknesses are important to consider when conducting any docking study. They do not undermine the utility of molecular docking methods and VS in drug discovery, but speak to the potential for improvement with the improvement of docking algorithms and scoring functions.

1.3.2.6 A comparison of popular docking programs

A review article in 2008 documented over 60 docking programs and 30 scoring functions, and many more have been developed to date [91, 92]. However, fewer than 10 are widely used [93]; these programs are listed in Table 1.1 with a general description of their sampling and scoring methods. AutoDock, Dock, and eHits are free for academic users, which greatly increases their usage within academic circles.

Each of these docking programs involves their own software packages, some with very powerful molecular viewers and modeling packages. For example, Glide and ICM both have a complete desktop modeling environment that can perform energy minimization, multiple sequence alignments, homology modeling, among other functionalities. In addition the methods of protein-preparation, ligand-preparation, docking settings, and input and output formats differ greatly between different programs. To this end, there is a learning curve to each program, and experts who know to fine-tune the details can obtain better docking results than novice or intermediate users [94].

With the abundance of docking programs, it is not surprising to find many studies comparing the speed and accuracy of these programs. 14 of these studies published from 2003-2009 have been summarized [95]. The gist of such studies is to select a set of 10-200 protein complexes (representing around 10 unique protein targets) and compare the docking programs based on their ability to: 1) dock the ligand in a correct conformation; 2) score the docked ligand reasonably; and/or 3) select the correct ligand(s) out of a library of decoy molecules. However, a major limitation to these studies is that they each use a different set of docking programs, and a different set of protein and ligand structures as a test set.

I review here the studies that assessed ICM, since it was the software used in this thesis. All the studies assessed whether the crystallographic pose of a PDB structure complex could be reproduced by docking within 2Å RMSD. Perola *et al*. studied over 200 PDB complexes for three protein targets, and found acceptable RMSDs in 61% of cases using Glide, 48% of cases using GOLD, and 45% using ICM [96]. Chen *et al*. tested GLIDE, GOLD, ICM, and FlexX on 12 proteins (164 complexes) and found that ICM performed best, correctly docking 91% of the complexes within 2Å RMSD [97]. GLIDE followed with 73%, GOLD with 55%, and FlexX with 43%. Bursulaya *et al*. chose a set of 37 protein complexes for 11 different proteins and found that docking passed the RMSD cut-off in 76% of cases for ICM, 46% for GOLD and AutoDock, 35% for FlexX, and 30% for DOCK [98]. Cross *et al*. performed one of the largest studies using 68 diverse structure complexes, and found the RMSDs acceptable in 72% of cases for ICM, 69% for Glide, ~55% for DOCK, ~50% for Surflex, and ~45% for FlexX [99]. This study also employed the Database of Useful Decoys (DUD) [100], a set of 40 diverse protein targets with known active compounds and physically similar but topologically dissimilar decoy compounds. When calculating the area under the curve (AUC) in receiver operating characteristic (ROC) curves, all methods performed better than random (AUC=0.5) with AUC values of 0.72 for Glide, 0.66 for Surflex, 0.63 for ICM, and 0.55 for DOCK. The Chen *et al*. study also assessed enrichment factors (EFs) - the ratio between the percentage of active compounds in the top X% scoring compounds, and the percentage of active compounds in the entire compound set. They found that at X=10 (selecting the top 10% scoring compounds as predicted binders), ICM had an EF of 6.1, Glide of 4.3, FlexX of 2.2, and GOLD of 1.71.

The overall consensus is that there is significant variability in VS results [99], as docking software performance greatly depends on the chosen target set and its molecular properties. However, the commercial software Glide and ICM consistently rank among the best performing software based on their docking and scoring accuracy. In contrast, the free software DOCK and AutoDock generally do not perform as well in assessment studies.

1.3.2.7 Validating docking results

With the low accuracy of docking predictions, experimental validation of docking predictions is essential. *In vitro* assays generally fall into two categories, target studies and cellular studies.

Target-based studies test for a direct binding interaction between the target and ligand. The BIACORE method fixes the protein to a surface and washes it with the drug, testing for changes in molecular weight to determine if the drug has bound to the protein [101]. Some methods measure the change in stability of a drug-bound protein, such as through a shift in melting point (thermal shift assays [102]), or through ease of proteolysis (the DARTS - drug affinity response target stability method [103]). Enzymatic assays test if the ligand can inhibit or increase the enzymatic activity of the target protein. An example would to use a kinase protein with a substrate peptide, radio-labeled ATP, and a predicted inhibitor, and seeing if adding the inhibitor reduces the output of radio-labelled peptide. The most convincing type of validation is solving a crystal structure of the target-ligand complex. However, it is also the most difficult to achieve from a time and cost standpoint.

Cell-based studies involve choosing a cell line for which the target is known to play an integral factor, or for which the target has been overexpressed. Treating cells with the predicted compound in increasing doses should then show the expected phenotype with increasing clarity. As a control, cells without the target aberration are used. For example, imatinib would kill CML cells but not control ML cells. Affinity pull-down methods fix the drug on a column, and wash through cell lysates. Proteins that bind to the drug can then be identified using liquid chromatography separation techniques or tandem mass spectrometry [43, 104]. Cell-based assays also have the potential to reveal the effect of the ligand on pathways through the use of antibody probes, and can give more insight into the molecular effect of the ligand. However, cell-based assays may need to be accompanied by biochemical binding assays to confirm whether an observed cellular effect is due to inhibition of the intended target or an off-target in the cell.

In vivo validation in animal models occurs after both the binding interaction and cellular effects have been verified. In the case of a tumor, a typical study would be to graft and grow the tumor in mice, treat the mice with the drug, and check for shrinkage of the tumor without compromising the weight of the mouse. A statistically significant difference in survival between test and control mice can be calculated using Kaplan-Meier survival curves.

Each method has its own caveats. In *in vitro* models, proteins or chemicals are often 'fixed' to beads or capsids or surfaces in regions that are supposedly not important for binding. However, this cannot be guaranteed. In cell-based assays the complexity increases with the involvement of the entire cell, and positive results such as cell death could be caused by any number of factors. This complexity only increases in *in vivo* models.

1.3.2.8 Virtual screening resources

Many resources exist for virtual screening. There are at least 14 publicly available databases and six commercial databases of chemical or bioactivity data (reviewed in [105]). Furthermore, natural product companies have their own extensive library sets that are available by request. Some of the most popular chemical libraries are listed in Table 1.2. The increasing size of these libraries underscore the need for docking algorithms to be fast – to virtually screen PubChem against one protein binding site using ICM at one minute per target would require 12,500 days on one processor, and nearly two weeks using a 1000-processor cluster.

1.4 Finding new targets for existing drugs

Nobel laureate Sir James Black once stated, "The most fruitful basis for the discovery of a new drug is to start with an old drug." Old drugs - approved drugs and candidate drugs that failed at a late stage in development - have already been optimized for their pharmacokinetic and safety/toxicity profiles. As such, a newly identified use for an existing drug could be evaluated directly in phase II clinical trials, and save 40% of the overall cost for developing a new drug. For failed drugs, modifications can be made to improve their efficacy for an off-target; since these chemical starting points have drug-like properties, they would be expected to have fewer toxicity issues compared to other approaches of drug design [110].

Reviews of the field indicate at least 24 existing drugs have already been repositioned for new therapeutic uses and another 17 are in various stages of development [15, 111].

1.4.1 Recent experimental efforts

There have been many innovative experimental efforts to systematically elucidate novel drug-target and drug-disease relationships in recent years. CombinatoRx's discovery platform phenotypically screened thousands of pairs of existing drugs to find novel therapeutic combinations [112]. The Connectivity Map study determined gene expression profiles for small molecule drugs and disease tumor and cell line samples, and linked together drugs that showed an opposite gene expression profile to a disease profile [113]. Iorio *et al*. searched for similar gene expression profiles between cell lines that were treated with drugs, and linked together drugs that may have similar mechanisms of action [114]. High-throughput kinase assays tested 38 kinase inhibitors 317 kinase proteins to determine their targets [51]. The SOSA approach (selective optimization of side activities of drug molecules) examines the off-target activities of known drugs, and uses SAR studies to modify the drugs in order to be potent towards a particular off-target [110]. For example, the monoamine oxidase-A (MOA) inhibitor minaprine was found to have a weak affinity for the muscaranic M1 receptor (Ki 17μM) and acetylcholinesterase (Ki 600μM). Minaprine was subsequently modified into two different compounds that no longer inhibited (MOA), one a nanomolar agonist of muscarnic M1 receptor, and the other a nanomolar inhibitor of acetylcholinesterase.

1.4.2 Recent computational efforts

Just as computational methods have been applied to screening and designing new drugs, similar methods are also being used to elucidate drugs that can be repositioned for novel therapeutic uses. Given the large number of druggable protein targets and existing drugs, it is infeasible to set up assays to test every single interaction in the laboratory. In addition to the time and cost required, a tailored activity assay must be developed for each protein and compound libraries of all existing drugs must be collated. I review here computational efforts to virtually assay drug-target interactions. There are two main types of studies: 'inverse'

docking studies that dock selected drugs of interest across many protein targets, and larger studies that search for interactions between many drugs and many proteins.

Ligand-protein inverse docking was first proposed by Chen and Zhi in 2001 [115, 116]. They docked 4H-tamoxifen and vitamin E to a collection of 2,700 PDB ligand-binding sites, and found that half of their top predicted targets of these drugs were implicated or confirmed by previous experiments [115]. Liu *et al*. inverse docked eight drugs across 1714 targets, and found that at 15 of 17 known targets of the drugs ranked within the top 20 scoring targets. Zahler *et al*. searched for novel kinase targets of three indirubin-3'-oxime derivatives within 327 kinase structure complexes, and validated PDK1 as a target with direct binding IC_{50} 1.5μM and low micromolar inhibition of the MCF-7 cell line [117]. The online docking server TarFisDock was developed to aid in inverse docking, using the software DOCK to dock a user's compound of interest to 698 protein structures [118].

For studies predicting novel drug-target interactions between many drugs and many proteins, similarity approaches have been the most common. Campillos *et al*. searched for drugs with similar side-effect phenotypes, hypothesizing that such drugs would be likely to share the same protein targets [119]. They predicted a network of 1018 side-effect relationships between 746 marketed drugs, and validated 13 chemically dissimilar drugs from different therapeutic indications to bind to same protein. Keiser *et al*. associated targets based on the similarity of their known binding ligands (the SEA or similarity ensemble approach), and tried to reposition drugs chemically similar to known ligands of the targets [120]. They were able to confirm 23 new drug-target associations, of which five had IC_{50}'s under 100nM. Of particular interest was a reverse transcriptase inhibitor that could be repositioned to the histamine receptor. Another approach is sequence order-independent profile-profile alignment (SOIPPA), which computes binding-site descriptors based on shape, physical properties, and evolutionary profiles of active-site residues, without regard for actual site residues [121]. Kinnings *et al*. used this method to search for human proteins with similar ligand binding sites to the bacterial InhA protein, and found the protein COMT [122]. The COMT inhibitor, entacapone (Comtan) was then showed to inhibit InhA with an IC_{50} of 80μM. Though the inhibition was weak, the safety profile of entacapone and its few side

effects suggested it to be a potential repositioning drug lead for tuberculosis. Finally, the IDMap resource predicts associations between chemicals and drug targets based on chemical similarity and co-occurrence of annotated bioactivities between chemicals [123].

Cross-docking is when a set of chemicals and a set of proteins are all computationally docked against one another. It is commonly used on small datasets as a benchmark to assess docking methodology. For instance, Huang *et al*. cross docked 40 protein structures to 40 ligands (native to the PDB structures) to assess ligand enrichment and specificity [100]. One interesting approach is 'in situ cross docking' where the proteins are placed side by side on one grid, and the ligand is docked to the multiple proteins simultaneously [124]. This study involved 3 proteins and 6 ligands, but whether the method is scalable to larger data sets remains to be seen. To date, the largest cross-docking undertaking has been the BioDrugScreen resource, which used AutoDock to dock 1592 diverse small molecules to 1926 binding sites on 1589 human targets [125]. This resource provides various types of scores for the docked interaction including AutoDock, GoldScore, ChemScore, and PMF score, for the user to judge.

1.4.3 Resources for drug-target interactions

There has been an explosion in new drug-target interaction databases developed in recent years, which reflects the increasing interest in drug polypharmacology and repositioning. Some of the popular databases are listed in Table 1.3. Many of these databases are manually curated, like DrugBank, and should have few false positive interactions. Two very recent databases PROMISCUOUS and ChemProt highlight the trend of attempting to integrate data from several sources into a comprehensive resource.

1.5 Thesis overview and chapter objectives

Human diseases are comprised of complex mechanisms involving aberrations in numerous proteins and pathways. We now know that small molecule drugs inhibit more target proteins than previously expected, and these off-target effects can contribute to drug efficacy. For approved drugs, finding novel off-target proteins can also lead to potential repositioning to other diseases or added insight into the drug's mechanism of action or adverse effects.

However, most successfully repositioned drugs to date have been discovered through serendipitous observations. Thus, the overall aim of this thesis was to develop computational methods to rationally predict drug repositioning candidates.

In Chapter 2, I describe an approach for performing large-scale molecular docking of many targets to many drugs to find novel drug-target interactions. The cross-docking of 4621 drugs to 2921 binding sites was a tremendous undertaking, requiring the use of a 1000-processor cluster. This is the largest cross-docking study and analysis to date. Due to the high false positive prediction rate of molecular docking, I developed a consensus-scoring threshold and combined it with rank information to retain only the most likely binding interactions. As a result, the predicted dataset was enriched for known interactions by over 50 times compared to standard docking software thresholds. The utility of this method was also confirmed when 31 of the top predictions that were not annotated in interaction databases were validated through literature search. Two approved drugs, one for asthma and one for cancer, were tested against the anti-inflammatory target MAPK14 and experimentally validated in *in vitro* kinase assays. In particular, the cancer drug nilotinib shows promise as a potential treatment for inflammatory diseases like rheumatoid arthritis.

In Chapter 3, I applied the docking approach to a particular drug target – the protein kinase Epidermal Growth Factor Receptor (EGFR). This target appeared to be ideal for docking studies due to PDB having over 30 solved crystal structures in complex with a wide variety of ligands. Known inhibitors of EGFR, selected as positive control ligands, scored and ranked highly using my method. When screening the DrugBank database against EGFR, the HIV pro-drug tenofovir disoproxil fumarate (TDF) appeared to be a potential repositioning candidate. I experimentally tested TDF in EGFR-overexpressing breast cancer cell lines, and found that TDF showed micromolar inhibition of these cancer cell lines as well as an inhibition of EGFR pathway signaling. However, subsequent *in vitro* kinase assays could not detect any inhibition of EGFR by TDF, suggesting that its effects on EGFR pathway signaling and cell proliferation may be mediated through other targets. This study underscored the challenging nature of drug-target interaction prediction, and the need to continue improving docking and scoring algorithms.

In Chapter 4, I applied the docking approach to screen 1,120 off-patent drugs against the triple-negative breast cancer (TNBC) target RSK. To improve the existing docking strategy, several approaches were pursued: building homology models to increase the number of useful RSK structures, defining new score and rank thresholds as well as visual criteria, and performing a high-throughput *in vitro* screen (HTS) in parallel. The docking methods greatly enriched for known RSK inhibitors compared to random. However the top 29 docking hits from the 1,120 off-patent drugs shared only six drugs in common with the top 32 HTS hits. Promising compounds from both the computational and experimental analyses were confirmed through secondary screens: a low-throughput *in vitro* binding assay, a western blot experiment, and a cellular proliferation assay. Three compounds inhibited RSK catalytic activity in a dose-dependent manner, blocked RSK signaling, and blocked TNBC cellular proliferation; these drugs represent repositioning candidates for TNBC, the most aggressive subtype of breast cancer with no targeted therapy options.

In Chapter 5, I present a personalized medicine case study, where the genome and transcriptome of a patient with a rare adenocarcinoma of the tongue was sequenced and analyzed. Discovering the RET pathway upregulation as a driving force of the cancer allowed us to rationally suggest two RET-inhibiting drugs approved for renal cell carcinoma. These drugs provided the patient, who had no standard treatment options, with 8 months of disease stabilization before the tumor metastasized. Sequencing the metastasized tissue and comparing it to the pre-treatment tumor revealed that an extensive amount of aberrations and evolution had accrued in the 8 months of drug treatment. I worked with many people to analyze candidate gene lists from copy number, expression, and mutation analyses, among others, to build a model of the disease mechanisms in the two tumor samples. This step formed the basis for suggesting rational therapeutic options for the patient. The success of using RET-inhibiting drugs to treat a tongue adenocarcinoma provided one of the first examples where finding genetic aberrations in the tumor allowed existing drugs to be repositioned for use in an individualized manner. This study also underscored the utility of next-generation sequencing methods for personalized diagnosis, by demonstrating the

effectiveness of drug repositioning for elucidated disease mechanisms without the need to discover novel drug-target interactions.

In addition to the work presented in this thesis, I have participated in other collaborative projects that are described in published manuscripts. With Dr. Fiona Brinkman and Debra Fulton, I helped develop a novel method of ortholog prediction [126] aimed at reducing the number of false positive predictions. I assisted Dr. Artem Cherkasov with assessing a novel method of calculating partial charges [127] as well as a novel docking strategy using QSAR models to improve docking speed [128]. With Dr. Cherkasov I was also involved in finding compounds that selectively targeted a Leishmania protein but not its human homolog [129, 130]. I have assisted Alexander Yakovenko with testing of a novel force field that incorporates another novel method of calculating partial charges [131].

Figure 1.1 A comparison of drug discovery pipelines.

a) The traditional drug discovery process where a novel drug is discovered through screening methods, optimized for maximum efficacy and minimal toxicity, and three phases of clinical trials before FDA approval. b) Drug repositioning starts with an existing drug so much of the discovery, optimization, toxicology, and clinical histories can be re-used, shortening the overall timeline and improving the chance of approval. Figure reprinted with permission from [15].

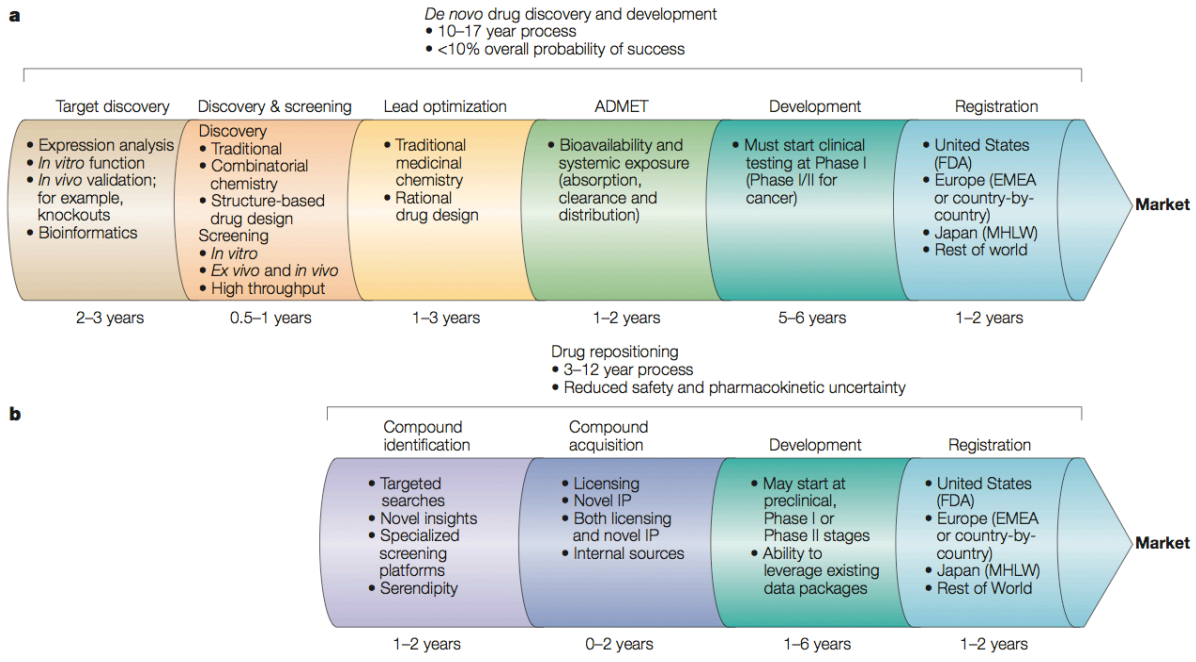


Figure 1.2 Global mapping of pharmacological space.

The human polypharmacology interaction network connecting drug targets (nodes) based on shared chemical binders (edges). There are 468 proteins in the network and 3,636 polypharmacology relationships. Reprinted with permission from [48]

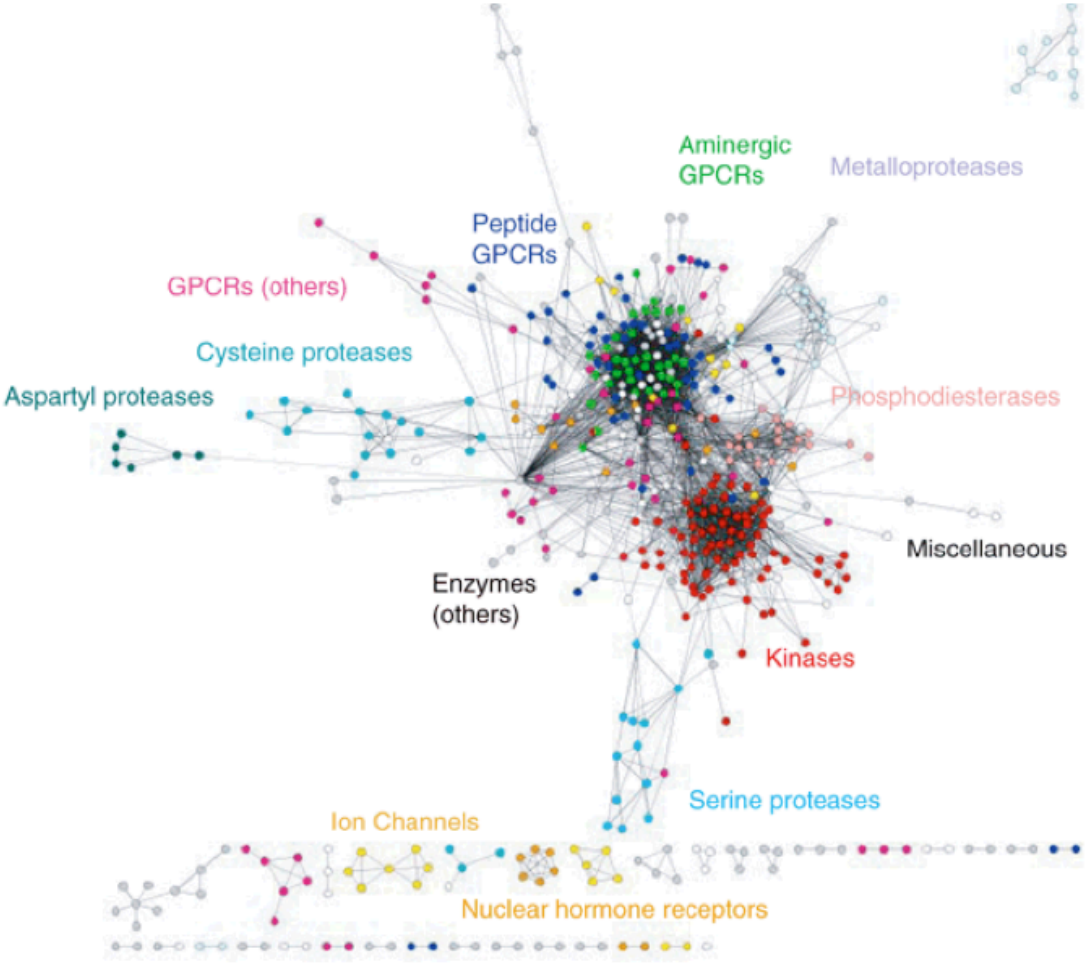


Figure 1.3 The main steps of a virtual screening procedure using molecular docking.

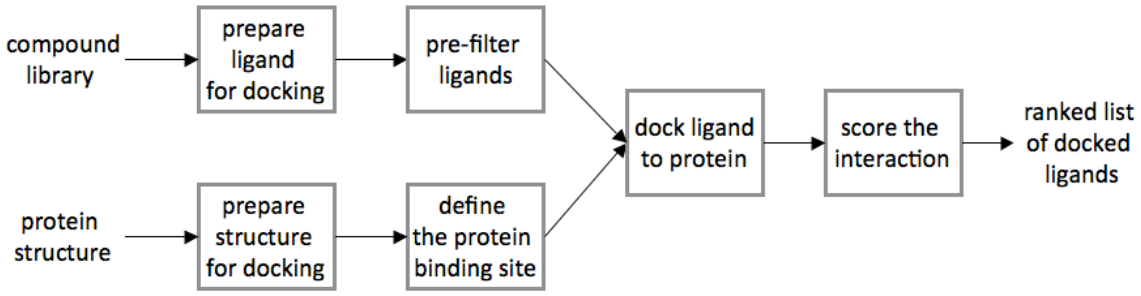


Figure 1.4 The ICM docking algorithm.

ICM uses a MC algorithm to sample the ligand conformational space in order to find the docked ligand conformation with a global minimum energy. At each step, a movement is applied to the ligand and a local energy minimization is performed to refine the ligand conformation. A more comprehensive energy is then calculated for the docked interaction. The Metropolis criterion is used to retain some poor scoring conformations to avoid falling into local minimum energies.

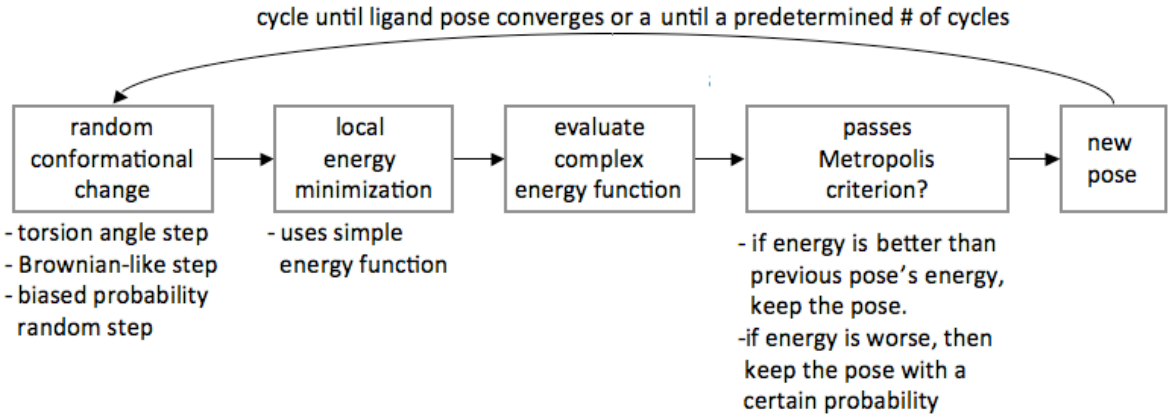


Table 1.1 A comparison of available docking programs.

Program	License type	Docking method	Scoring method
AutoDock 4.0 [132]	Academic	Genetic algorithm	Force-field / Empirical
Dock 6.1 [84]	Academic	Multiconformers & Incremental construction	Force-field
eHits [133]	Academic	Incremental construction	Empirical
FlexX [134]	Commercial	Incremental construction	Empirical
Gold [135]	Commercial	Genetic algorithm	Force field
Glide [136]	Commercial	Stochastic search	Empirical
ICM [87]	Commercial	Monte-Carlo	Empirical
Surflex [83]	Commercial	Incremental construction	Empirical

Table 1.2 Popular chemical compound libraries for virtual screening.

Library	Description
NCI 3D [137]	A public repository of about 200,000 physically available compounds.
ZINC [138]	2.7 million commercially-available compounds in docking-compatible format
ChemDB [139]	5 million commercially available small molecules that can be synthetic building blocks
Pubchem [140]	A public repository of over 18 million small molecule compounds
GDB-13 [141]	970 million virtually generated, chemically possible, drug-like organic compounds up to 13 atoms in size

Table 1.3 Popular drug-target interaction databases.

Library	Description
Pubchem Bioassays [142]	National Institute of Health (NIH) repository of over 2300 screening studies [143], containing the biological activities of small molecules.
DrugBank [144]	A fully curated database of 6826 drugs, with information about their mechanisms, pharmacology, and protein targets.
KEGG Drug [145]	Known target protein and target pathway information for over 3000 approved drugs in Japan, USA, and Europe
WOMBAT [146]	A curated database with small molecule biological activity information for more than 1400 protein targets.
MDL Drug Data Report (MDDR) [147]	A commercial database of biological activities of over 100,000 chemicals, compiled from patent literature [148]
PROMISCUOUS [149]	A database integrating DrugBank, SuperTarget, and SuperCyp data, focusing on drug-target interactions, and side effect information. They also use text-mining to find interactions.
ChemProt [150]	A database integrating ChEMBL, BindingDB, DrugBank, PharmGKB, WOMBAT, and PubChem Bioassay, CTD, and STITCH. It includes drug-protein associations between 700,000 chemicals and 30,578 proteins that may not be direct binding events.

2 A Large-scale Computational Approach to Finding Novel Targets for Existing Drugs

2.1 Introduction

Drug repositioning is the process of finding new therapeutic indications for existing drugs. It is an efficient parallel approach to drug discovery, as existing drugs already have extensive clinical history and toxicology information. However, novel interactions between drugs and target proteins must first be discovered. This is not a simple task – there are at least 30 approved drugs with unknown cellular targets and mechanism of action.

Drug candidates are routinely screened against a small panel of similar proteins to determine their specificity to the intended target. Large panels with hundreds of kinase proteins have been developed to assess kinase inhibitor specificity [51], especially since we now know that many kinase drugs are multi-targeting. However, the druggable proteome is much larger than just the kinome, so larger and more varied protein panels are needed to truly assess drug specificity. With the availability of massively parallel DNA sequencing technology, recurrently mutated proteins in diseases – such as EZH2 in certain lymphomas [151] and FOXL2 in certain ovarian cancers [152] - are now being rapidly determined and are also relevant drug targets. However, testing all drugs against all targets experimentally is extremely costly and technically infeasible.

Recent computational endeavors to predict novel drug repositioning candidates have used methods incorporating protein structural similarity [122, 153], chemical similarity [120], and side effect similarity [119]. Molecular docking is a computational method that predicts how two molecules interact with each other in 3-dimensional space. It is well established as a virtual screening method in drug discovery [52], where typically many chemicals are docked against a specific protein binding site, in order to discover novel inhibitors of that target. Compared to similarity analyses, docking has the potential to find drugs that bind to proteins with novel scaffolds as well as off-targets that may be structurally dissimilar to the known

targets. Recently, Brylinski *et al*. used a machine learning approach combining protein sequence, structure, and ligand docking similarities [154].

Inverse docking is a method where specific ligands are docked to a large protein database to virtually screen targets instead of compounds. It was first used to predict potential off-target interactions and toxicities of 4H-tamoxifen and vitamin E [116]. Rockey *et al*. have inverse docked three kinase drugs to kinase crystal structures and homology models [155]. This approach has also been used to filter interactions predicted through protein binding site similarity [153]. More recently, the DOCK program was used to search for interactions between 10 Alzheimer's drugs and 401 proteins [156].

Large-scale docking of many targets to many drugs is now feasible when run on powerful computer clusters. However, limitations in scoring methods result in high false positive prediction rates [157], and large-scale studies amplify these low prediction accuracies. For example, the BioDrugScreen resource, which performed large-scale cross-docking using AutoDock, did not provide any methods to interpret results [125]. Here I present a molecular docking analysis of 4621 known drugs against 252 known protein drug targets for the prediction of novel drug-target interactions. This method emphasizes removing false positive predictions using protein structures determined to be reliable for docking, as well as consensus scoring and ranking thresholds. In short, I sought to retain only the highest confidence interactions as drug repositioning candidates.

2.2 Results

2.2.1 Computational pipeline

A computational pipeline was developed for large-scale molecular docking of drugs to protein targets (Figure 2.1). Briefly, I collected all 3D structures available for each drug target, determined binding pockets in the structures, and docked drugs to each pocket. Results were collected and thresholds were applied to select the top predicted interactions, which were then visually inspected.

2.2.2 Known drug-target interaction docking

I first docked 3570 known protein-drug interactions annotated by DrugBank, between 678 unique human proteins and 1309 small molecule drugs. I used the ICM docking program developed by Molsoft [87], which ranks ligands using a Monte-Carlo based docking procedure and an empirical, energetics-based docking score. Like most docking software, ICM recommends a standard score cut-off for virtual screening efforts: -32 [90], where more negative scores represent more likely binding interactions. However, studies have used different cut-offs (i.e. -28 [158]) depending on the protein target. Here I used a score of -30 as the threshold for 'good' dockings scores. Of the 3570 known interactions docked, 1116 (31%) had a good ICM docking score. 252 proteins had at least one known interaction predicted by docking – these formed the 'reliable' set of proteins that are hypothesized to be more suited for docking purposes. A breakdown of protein classifications for this reliable set revealed that 67% of targets were enzymes, of which 12% were protein kinases. In contrast, there were few G-protein coupled receptors in the database due to lack of crystal structures, which reflects both the current state of solved protein crystal structure space as well as popular drug targets.

2.2.3 Known drug-target interaction docking evaluation

In high-throughput molecular docking, it is common to hold protein structures rigid during the simulation. With this restriction, re-docking a PDB ligand back to its native PDB structure (cognate docking) is a simpler task than docking a different ligand to the structure (non-cognate docking) because in the former case the protein is already in a specific ligand-bound conformation. Due to the abundance of existing protein-ligand complex structures in PDB, cognate-docking situations occur frequently. Previous studies show that such cases can be docked well in 60-80% of cases [159]. In contrast, the more informative non-cognate docking is only successful in 20-40% of cases [159].

The 1116 known interactions were examined as to whether those that docked well were only due to docking cognate ligands. For each interaction, I observed whether the drug bound 1) a *holo* (unliganded) protein structure, 2) an *apo* (liganded) structure with a same or similar ligand as the drug (the cognate-docking scenario), or 3) an *apo* structure with a chemically

different ligand from the drug. Chemical similarity was defined as having a Tanimoto coefficient less than 0.54. Figure 2.2 shows that cognate docking occurred in 380 of the 1116 interactions. Of these, only 56 were drugs docked to an *apo* protein with the same ligand (Tanimoto coefficient of 0). The majority of drugs docked well to *holo* structures as well as *apo* structures with dissimilar ligands. In short, the ICM docking method was able to predict known interactions for both cognate and non-cognate docking scenarios.

Aside from the docking score, it was also important to verify that the ligands were docked in correct binding conformations. Further examination of the 380 cognate dockings revealed that the docked drug conformation was close to the known drug conformation (RMSD value ≤ 2Å) in 69% of cases. The other 31% fell into two categories: 1) partly symmetrical ligands like NAD and 2) ligands that bound to a small pocket. In the first case, the molecule was incorrectly determined to be flipped, causing a high RMSD; however, its central portion was docked correctly due to symmetry. In the second case, the region of ligand bound in the pocket was docked correctly, but the region free in solvent contributed to a poor RMSD value. Overall, this analysis showed that when a known interaction was docked with a good score, the binding conformation was also reasonably predicted.

2.2.4 Known drug-target interaction network

I gathered the known protein-drug interactions into a network (Figure 2.3) with proteins as rectangular nodes, drugs as circular nodes, and interactions as edges. Interaction edges with good docking scores were highlighted in red. Proteins from the same family were often grouped close together and shared many drug interactions, such as the retinoid X and retinoic acid receptors and the matrix metalloproteinases (RXRs, RARs, MMPs). Proteins having the most known drug interactions in the network included the transport protein serum albumin (ALB) and the phosphatase PTPN1. The most highly connected chemicals in the network were metabolites: ATP, NAD, and NADP. For some proteins such as MAPK14, 13 of 14 known inhibitors were well predicted by docking, whereas for others such as the angiotensin-converting enzyme (ACE), only one of its nine known inhibitors scored well. For 426 of the 678 protein targets not included in Figure 2.3, none of their known interacting drugs could be docked well, reflecting the limitations of current molecular docking methods. To this end, I

chose the subset of 252 protein targets for which at least one known drug docked well (from the 1116 interactions that docked well), which were deemed as more 'reliable-for-docking' compared to the other proteins.

2.2.5 Large scale cross-docking and score thresholds

I proceeded to dock the 252 reliable protein set against the database of 4621 drugs. Considering the multiple crystal structures per protein and the multiple binding pockets per structure, there were a total of 1514 crystal structures and 2923 binding pockets. Each drug was docked to all binding pockets of a protein and whichever pocket gave the best docking score for the drug determined the final protein-drug score. This method allowed multiple conformations of a protein to be accounted for during docking and provided a simple model of protein flexibility.

In total, 1.2 million protein-drug interactions were docked. 104,625 (0.9%) had ICM docking scores (icm-score) of -30 or better, encompassing all 1116 known interactions in the reliable data set. Since the fraction of known interactions in the predicted set was so low, I assumed that the vast majority of predictions were false positives. Though I believed that novel drug-target interactions existed and were enriched within these 104,625, there was clearly a need for more stringent score thresholds.

2.2.6 Investigating score thresholds

Various methods of selecting top drug-target interactions were investigated. The standard software-recommended icm-score is based on a weighted sum of various binding energy terms [87]. The pmf-score, or potential of mean force score, is a measure of the statistical probability for the drug and protein to interact with each other (for example, it examines interatomic distances and atom types of the docked interaction and compares that to existing distances in PDB) [90]. A consensus score was developed that uses both icm- and pmf-scores and allows us to select the top X% of interactions for each protein; it is described in more detail in case studies below. Interactions were ranked in two ways. The drug-rank is the rank of this drug compared to all drugs docked to this protein (from 1-4621), and the protein-rank is the rank of this protein when the drug is docked to all proteins (from 1-252).

Requiring high drug and protein ranks (i.e. a low value when the two ranks are summed together) enforces a mutual specificity criterion. I hypothesized that by choosing interactions with good scores and ranks, more false positive predictions would be filtered out.

The positive predictive value (PPV), defined as the proportion of predicted interactions that are known binding interactions, was measured to assess performance. The premise is that a better threshold would yield a set of predictions more enriched with known interactions as well as novel interactions that are more likely be true binding events. Figure 2.4a shows that as the stringency of a threshold increased (i.e. icm-score of -40 versus -30), fewer interactions were predicted; however, the PPV increased due to a higher proportion of known interactions in the predicted set. This behavior was consistent for all thresholds, and the highest PPVs were generally observed within the top 100 predicted interactions. It is important to note that each of the 4621 drugs will always have a top-ranked protein (interactions with protein-rank of 1), and each of the 252 proteins will always have a top-ranked drug (interactions of drug-rank 1). Thus, the protein-rank threshold particularly was not sensitive alone.

The protein-rank and pmf-score thresholds appeared to be the worst based on both the PPV plot (Figure 2.4) and on enrichment factors (Table 2.1). However, they showed better PPVs when combined with other thresholds. For example, the drug rank and protein rank measure performed much better than drug-rank alone, and the consensus score (combining icm- and pmf-score) also performed better than the icm-score alone. When the enrichment factor was measured for each type of threshold at its most stringent setting (leftmost points of Figure 2.4a), the pmf-score and protein-rank were the least effective at predicting known drugs (Table 2.1). Instead, combinations of score and rank criteria provided a 100-500 fold enrichment of known interactions compared to a random algorithm, and 10-50 fold enrichment compared to a standard binding energy-based ICM score cut-off of -30. Interestingly, the drug-rank 1 and protein-rank 1 combination threshold (basically a sum of ranks of 2) performed surprisingly well; however, adding the consensus score clearly improved the PPV for the top ~300 interactions (Figure 2.4b) which were the most interesting to us for further inspection.

Another threshold method is to use the scores of known binders as the score cut-off for each protein. Table 2.2 shows that using the best and worst icm- and pmf-scores of known drugs did not result in a higher enrichment, nor did it help narrow down the number of predicted interactions.

Overall, the combination of consensus score with the two ranks gave the highest PPV and enrichment values: in the top 50 predicted interactions, 49% were known. This gave us confidence that many of the other 51%, all novel interactions, were real.

2.2.7 Case study: MAPK14

Two examples are presented to illustrate the utility of combining rank and scoring criteria. The first is for the signaling protein MAPK14 (also known as p38 alpha), an integral component in numerous cellular processes. It is a drug-target for inflammatory diseases [160]. MAPK14 is known to be a challenging docking target due to its structural flexibility [161] and its shallow binding pocket [100]. However, these docking studies used only one 3D structure. In my dataset, there are 35 crystal structures of MAPK14 in different conformations, providing a simple view of protein flexibility.

2.2.7.1 MAPK14 docking results and consensus score threshold

The consensus score is based on the observation that when docking a large number of diverse compounds to any target, most compounds have poor icm- or pmf- scores, and few compounds have both good icm- and pmf- scores. Therefore, I chose a linear threshold that eliminated the densest area of points in the poor scoring region (top-right) of a score plot like Figure 2.5, and thus selects the compounds in the best scoring region (bottom-left) as potential interaction hits. In theory, the consensus score picks drugs with good icm-score and pmf-score ranks. As seen in Figure 2.4a and Table 2.1, the consensus score performed better for PPVs and enrichments compared to a simple icm- and pmf- score combination.

Figure 2.5 plots the icm- versus pmf- scores of the 4621 drugs docked to MAPK14. Each drug is a point on the graph, where the 5% of drugs passing a consensus threshold are shown

in orange, and the 1% passing a consensus threshold are shown in purple. For 67 drugs, MAPK14 was one of the top 5 scoring targets; they are circled in green. Table 2.3 shows that a combination of the consensus and protein rank criteria resulted in the best enrichment (110x) of known drugs. There were 15 annotated known binders of MAPK14 in DrugBank, but I disregarded 2-chlorophenyl due to it being a very small molecule with a very weak MAPK14-binding affinity (>1mM). 10 of 14 known drugs were predicted through the stringent thresholds. Though 4 true positive binders were lost, 99.99% of points were eliminated, presumably consisting mostly of non-binders. Through literature search, it was found that imatinib and quercetin have been previously tested against MAPK14 and did not show any inhibition [162]. This suggested that the 5% consensus threshold was too lenient for MAPK14, whereas the 1% was more appropriate. Within the other approved drugs predicted to bind MAPK14, literature validation was found for sorafenib, a multi-kinase inhibitor approved for renal cell carcinoma [163], and gefitinib, a EGFR inhibitor approved for late stage non-small cell lung cancer [43].

2.2.7.2 Experimental validation of two MAPK14 predicted inhibitors

Previous high-throughput studies have shown varying results regarding nilotinib-MAPK14 inhibition. Some enzymatic assays to MAPK14 showed weak inhibition: 570nM or 2.2μM depending on the assay type [164]. Direct binding assays have shown 100nM Kd [164] or no binding at all in a peptide pulldown experiment [162]. Since nilotinib was one of the top approved drugs predicted to bind MAPK14, I decided to further experimentally validate the interaction. MAPK14 ATP-competitive binding assays were performed for two inhibitors that were available for purchase: zafirlukast, and nilotinib. As seen in Figure 2.6, both drugs exhibited inhibition of MAPK14 at therapeutically relevant concentrations (<10μM) in a dose dependent manner. Zafirlukast (AstraZeneca) is an oral leukotriene inhibitor that reduces inflammation of the breathing passage in asthma patients. I found that it does inhibit MAPK14 weakly, and this may contribute to its inflammation reducing effect. The chronic myeloid leukemia drug nilotinib was especially potent with an IC_{50} of 40nM.

Despite their appeal as an inflammatory disease target, MAPK14 drug candidates to date have failed due to drug toxicity issues [164]. Though it may seem underwhelming to use a

cancer drug with potentially serious side effects to treat inflammation, nilotinib is noted to have a much milder adverse effects profile compared to its similar drug dasatinib [162]. Another similar drug imatinib has shown promise in treating rheumatoid arthritis in mouse models [165] and specific patients [166, 167], speculated due to its inhibition of mast cell c-Kit and PDGFRB. Nilotinib also inhibits these two proteins, and its extra inhibition of MAPK14 may render it a better choice for arthritis models. Recently, nilotinib was tested in a glucose-6-phosphate-isomerase-induced arthritis mouse model and found to significantly prevent paw inflammation – to a greater extent than imatinib [168]. This study also suggested that the two drugs acted through some distinct mechanisms. Overall, these findings seem to agree with the observation that nilotinib potently inhibits MAPK14, unlike imatinib, and thus has added potential as an anti-inflammatory drug.

2.2.8 Case study: BIM-8

A second example is the Protein Kinase C inhibitor BIM-8. I docked BIM-8 to the set of 252 reliable targets, and the results are plotted in Figure 2.7. Each point on the graph represents a protein target, and targets for which BIM-8 passes the 5% consensus threshold are shown in orange.

These results were compared to three previous studies. Two studies performed protein kinase assays with radioactive ATP and substrate peptides, where inhibitor binding decreased the amount of radioactive peptide produced [169, 170]. The third study performed thermal shift assays where inhibitor binding increased the kinase stability and thus the melting point [171]. BIM-8 targets discovered by these papers are shown in shades of red in Figure 2.7, and non-binders in these papers are shown in green. The only annotated target of BIM-8 in DrugBank is PDPK1. GSK3B and PIM1, which are in the top 5 protein rank and top 5% consensus threshold, were also validated as inhibitors. PDPK1 was not found to be an inhibitor by the first two studies but was confirmed as a binder by the third study with a kinase assay and crystal structure. Overall, there were 4 known binders (PIM1, PDPK1, GSK3B, LCK, since CDK and MAPK14 are probably weak or non-binders) and found that applying a 5% consensus threshold and protein rank criteria gave us 63-fold enrichment over random

selection, and a 63/10.5= 46 fold enrichment over a standard ICM score threshold of -30 (Table 2.4).

2.2.9 Drug-target interaction map

For a global and quantitative review of the predicted protein-drug interactions, I plotted the icm scores of drugs docked to established drug targets (Figure 2.8). Each protein is represented by a row, on which a black cross denotes a known drug docked to the target, a red dot denotes an approved drug docked to the target, and a blue dot denotes an experimental drug docked to the target. Only protein-drug interactions that docked with a score passing the consensus threshold and had a protein-rank ≤ 5 are shown.

Overall, the known drugs (black crosses) had better scores than other drugs for a given target. This was expected, as many of these known drugs were chemically optimized for their targets. For a number of targets, the known drug was the only predicted interaction. None of the approved and experimental drugs from DrugBank were able to dock well, despite a reliable protein structure, suggesting that virtually screening larger chemical databases may be the only way to discover novel inhibitors by docking. For most targets, at least one experimental drug showed a better score than the known drugs; however, experimental drugs are often unavailable for purchase or experimental testing. Instead, I was most interested in cases with approved drugs such as the MAPK14-sorafenib example which was verified by the literature, and the MAPK14-nilotinib example which was verified with an *in vitro* kinase assay.

Through literature search, I found experimental support for many of the top drug-target predictions that scored better than known interactions (Table 2.5). These all pass the 1% consensus threshold and are observed to have high drug and protein ranks for the most part. It is important to note that the drug-rank depends on the number of known binders for the protein; thus, since ESR1 had 39 annotated drugs in DrugBank, a drug-rank of 16 is not low. In contrast, a drug-rank of 16 would be low for MMP13, which has only seven annotated drugs in DrugBank.

One type of validated interaction includes drugs that are close analogs of known drugs for that target; for example, the estrogen analog ERA-923 is a known selective estrogen receptor modular (SERM) [172]. Genistein is known to bind both ESR1 and ESR2 [173]. Becocalcidol and ED-71 are vitamin-D analogs and bind the vitamin D receptor [174, 175]. Drosiprenone is a synthetic progestin with anti-mineralocorticoid receptor (MR, NR3C2) effects and has potential for reducing cardiovascular risk in women taking oral contraceptives or postmenopausal hormone treatment [176]. Another type includes interactions that were missed or mistyped by drug-target databases in the manual curation process. BMS181156, for example, is known to bind RARG and has even been solved in a X-ray structure complex but is entered in DrugBank 2.5 as binding only to RARG2. The interaction has been corrected in the latest version of DrugBank, however this example highlights the utility for my computational method to assist in curating drug-target databases.

Due to the many in depth studies on kinase inhibitor specificity, collaborating evidence was available for some of the kinase protein interaction predictions. For example, vatalanib is a known pan-VEGFR inhibitor [177], nilotinib is a potent KIT inhibitor [178], and other inhibitors of MAPK14 and targets of kinase inhibitor BIM-8 were discussed in previous sections. The subset of predictions for which there was literature support included many structural analogs. I performed a similarity analysis for the top 45 interactions at the strictest threshold (Table 2.1, bottom row), where each drug was compared to all known binders of its target. For the 22 known interactions in the set, all drugs were similar as expected (within a threshold Tanimoto coefficient of 0.54). For the 23 novel predicted interactions, 14 were similar to at least one of the known binders and 9 were different. Thus, the docking approach at strictest thresholds was also capable of predicting non-structural analogs of existing drugs.

Overall, I was able to find literature support for 31 top interactions, validating the utility of my computational method for finding novel drug-target interactions.

2.3 Discussion

The binding of a small molecule drug to its target protein in a cell is much more complex than a single docking calculation. For example, an ATP-competitive kinase drug would have

hundreds of ATP-binding sites to choose from due to the large size of the kinome. Cancer drugs such as sunitinib are now known to potently inhibit many more kinase targets than previously expected [42]. In addition, non-kinase targets of kinase drugs have also been found: NQO2 was the first non-kinase target discovered for imatinib [162, 181], and several cytotoxic LIM kinase inhibitors were found to be actually inhibiting tubulin [182]. Such studies imply that the target search space for any inhibitor should be the entire druggable proteome.

The overall strategy was to find novel drug targets of existing drugs by computationally screening the druggable proteome. For this purpose, I chose molecular docking due to its speed, low cost, and detailed three-dimensional simulation. Moreover, docking can evaluate any protein with a solved structure due to its virtual nature, without the need for tailoring enzymatic assays or collecting drugs in solutions. However, docking is known to have a high false positive prediction rate, due to limitations such as incomplete binding pocket prediction, inadequate ligand conformation sampling, inaccurate scoring functions, lack of protein flexibility, and lack of water and cofactor molecules during the simulation. As evidenced in this study, only 31% of the 3570 known interactions docked with a good score. One review states that 10-50% of a set of diverse compounds can be expected to be docked correctly for a given target [157]. My results were well within this range, and I believe the method performed quite well considering the large variety protein targets involved and the automated nature of the pipeline. However, the other 69% of known interactions were not predicted due to docking limitations.

The computational method attempted to address these limitations. First, I manually included binding pockets that were present in PDB structure complexes but not predicted by the binding pocket search. Second, I docked each interaction 10 times to better sample ligand conformations. Third, I applied consensus score and rank criteria to further narrow down top scoring docking hits. Fourth, I used all available structures of a protein (versus choosing one representative structure), to allow a simple view of protein flexibility. I did not incorporate water and cofactor molecules in the docking simulations due to the computational complexity involved. However, by selecting proteins for which at least one known drug docked and

scored well, I obtained proteins for which the limitations of the docking software were not critical for a good prediction. In short, assuming the docked conformation of the known ligand was correct, I used only proteins for which the binding pocket was genuine, the scoring functions were adequate, the protein was in a conformation amenable for drug inhibition, and the lack of water or cofactor molecules didn't drastically affect the prediction.

Virtual screening studies typically involve docking large chemical databases to one protein target, selecting compounds that score within the top 0.5-1% of the database and then further prioritizing them by visual examination. When experimentally validating these top candidates, a 5% hit rate can be considered a successful endeavor (where a good hit is a predicted compound showing an experimental binding affinity in the μM or lower range) [183]. Depending on the target, the crystal structure, the software used, post-docking criteria (such as chemical clustering), and even the individual performing the visual examination, the hit rate can be improved to 10-40% (Cavasotto *et al*. had 14% hit rate from 50 tested compounds [158]; Sabio *et al*. had a 36% hit rate from 56 tested compounds [184]).

In this case, both the standard scoring threshold and the known-inhibitor score were not sufficient. With a normal score threshold of -30, docking 4621 drugs against 252 proteins resulted in 104,625 predicted interactions. This is roughly 1% of the docked interactions, so even selecting the top 1% of the docking hits for validation becomes prohibitive for large-scale studies. It is important to note that each protein has different physiochemical properties: for some proteins, hundreds of compounds pass the -30 cut-off, while for other proteins none pass. Thus, using the known-inhibitor score as a cut-off allows for a threshold that is tailored to each protein. However, this method still predicted ~8000 interactions at the most stringent. The consensus threshold allowed us to pick the top 1% (or any X%) of docked compounds with the best icm- and pmf- scores for each protein and further filter from there. After testing many combinations, the consensus score with rank information resulted in the highest PPV – nearly 50% - and enrichment factor – 50 times better than standard -30 score threshold and 490 times better than random selection. This high enrichment for known interactions suggests that many of the other predictions that have not yet been experimentally tested may be true binding interactions.

There are limitations to this scoring scheme. Since the pmf-score is a statistical score comparing the docked interaction to known interactions in PDB, a chemical with a different scaffold or novel binding conformation may have a poor pmf-score and become predicted as a false negative. However, my foremost goal in this study was to eliminate as many false positive predictions as possible and obtain a high enrichment of true positives in the predicted interaction set. Thus, it was acceptable to miss some false negative predictions. In addition, the consensus score is a simple linear separation method and may not be as powerful as a machine-learning algorithm that trains on known ligand docking scores. However, training algorithms require many known binders and non-binders for each protein, which is not the case for many proteins in the data set. In addition, the trained system may not be representative for proteins not in the training set or having few known binders. I desired an automated scoring method that did not depend upon the existence of known ligands. That is, if a protein structure had just one, or no known binders, the method would still be able to select the top 1% of docking hits.

Despite the limitations, my computational method has several novel points of interest. First, the 'reliable-for-docking' target set was based on whether any of a protein's known inhibitors could be docked well. This differs from most methods to date which do not restrict targets or use only protein structures in complex with a ligand [156]. It was shown through the known interactions analysis that many non-cognate dockings could also be successful. Second, the consensus score line differs from most methods to date which are simple score thresholds [158] or utilize machine learning algorithms [185-187]. I observed for every target, that most docked drugs have poor icm- and pmf-scores and are clustered densely on the top-right of a score plot. The consensus score line thus chooses the top X% of drugs by removing the N-X% (where N is the total number of drugs) in the smallest possible top-right area. Recently, Yang *et al*. applied 2-directional normalization of docking score matrices to select top interactions, which in essence selects drugs that have both a high drug-rank and protein-rank [156]. For this dataset the protein-rank and drug-rank combination performed well, but that adding the consensus score allowed for even higher PPVs (Figure 2.4). The former method may predict more false-positives when an interaction has high ranks but poor scores – such

as when there are no real interactions to be found for that drug and that target. I thus believe that including the docking scores into the threshold is important.

To date, cross-docking of proteins to compounds has generally been used for small datasets. As an example, Huang *et al*. docked 40 targets against 40 compounds to check whether their docking method could distinguish between a target's cognate ligands and the other targets' cognate ligands [100]. In this large-scale cross-docking study, the use of a 1000-processor cluster was essential to completing the tens of millions of docking simulations in a timely manner. In addition, the large number of crystal structures and binding pockets involved required much of the docking pipeline be automated.

High-throughput computational screening of drug-target interactions represents a parallel approach to high-throughput experimental screening. Due to differences in experimental methods, assay settings, and protein panels, different studies may present differing results. For example, small molecule affinity purification methods that use whole cell lysates would give different results from *in vitro* kinase assays that use a specific panel of proteins. In the case of gefitinib, two such studies had distinct differences in their proposed cellular targets [42, 43]. Differences in methods are also further compared in a study by Manley et al [164]. I presented an example for BIM-8, which binds to PDPK1 differently in two similar *in vitro* experiments. For MAPK14, the experimental results for nilotinib also varied. Two purchasable approved drugs were experimentally tested against MAPK14; nilotinib was a strong nanomolar inhibitor, and zafirlukast was also an inhibitor, though not as potent. Thus, interactions that are predicted to be very likely inhibitors computationally may merit extra study even if experimental tests are initially negative.

In short, I have developed a computational pipeline that can run large-scale cross-docking of compounds to targets. I developed stringent criteria to filter a large proportion of false positive interactions. The two case studies presented were selected based on known experimental binding assay data, so as to demonstrate the notable enrichment of known interactions using the scoring and ranking criteria. I hypothesized that predicting a set of interactions with a higher PPV (enrichment of known interactions) would also lend

confidence to the other novel interactions in the set. This appears to have worked, as I was able to find validation for 31 predicted drug-target interactions that were not previously annotated in DrugBank, as well as validate two other inhibitors of MAPK14. Other drug-target interaction predictions are currently undergoing experimental validation; novel interactions discovered are potential drug repositioning candidates, but also provide insight into a drug's mechanism of action and adverse effects profile.

2.4 Methods

2.4.1 Pocket database and drug database construction

The DrugBank 2.5 database [144], containing drug information and comprehensive information of their targets, was downloaded. I extracted human protein drug targets from DrugBank and retrieved their sequences from SwissProt [188]. Protein Data Bank structures showing at least 95% sequence identity for proteins at least 20 amino acids in size were downloaded. They were required to be X-ray crystal structures with a minimum resolution of 2.8Å. Multiple chains were grouped into a set of non-redundant sequences, based on PDB's chain redundancy analysis at the 95% sequence identity level.

2.4.2 Preparing a target pocket database

Protein structures were prepared for docking using Molsoft's ICM software version 3.4-9c [87], removing water molecules, solvent ions, and other ligands from the structures. I added hydrogen atoms to the structures then optimized their positions.

To predict pockets, or potential binding sites, I used the PocketFinder [78] method in ICM, which calculates a Lennard-Jones transformation of the van der Waals energy for an aliphatic carbon probe on a grid map. The grid potential values are smoothed over 10 iterative averagings. Contouring of the map defines the binding pockets, and those with volume under 100 $Å^3$ are removed. For each protein structure, the three largest pockets are retained in the database. If metal ions were found near a pocket, two receptors were prepared for docking: one of the protein with the metal ion and one without.

The receptor was defined as the box 3.5Å surrounding the pocket. If the pocket overlapped well with the ligand but the ligand extended out of the protein structure, the receptor was defined to be the box 3.5Å around the pocket but also including 2.0Å around the ligand. This ensured that known ligand binding sites not predicted by my automated method were also included in the pocket database.

2.4.3 Molecular docking procedure

Drugs were docked to target receptors using the ICM virtual library screening (VLS) module. This method performs rigid-receptor flexible-ligand docking using a two-step Monte Carlo minimization method and energy scoring function to sample ligand conformations and select the best docking hits. MMFF partial charges and ECEPP/3 force-field parameters are used. ICM virtual screening automatically assigns the ligand ionization (charging carboxylate, phosphate groups), stereochemistry and tautomeric forms. I did not include alternative ligand forms, as this would have caused too much increase in the size of the chemical database.

Docking one interaction required on average 30 seconds to 1 min per processor. A given protein may have several structures, each of which with more than one pocket; in such cases I dock all pockets to a drug, and the best scoring interaction is selected to be the representative protein-drug score.

To ensure a sufficient coverage of the docking energy landscape, each drug-target interaction was docked 10 times in the known docking analysis and 5 times in the large-scale cross-docking analysis. Docking was performed on a Linux cluster with 1000 processors – this level of throughput allowed us to complete 1-3 million dockings per day.

2.4.4 Known interactions docking

8867 known interactions between human protein targets and drugs were culled from the DrugBank Drugcards database. Of these, 3570 interactions with protein target crystal structures present in the database were docked. Due to the Monte-Carlo nature of the ICM

method, each interaction was docked 10 times to better cover the docking energy landscape. After 10 iterations, the best scoring prediction was retained.

If the protein structure was solved in complex with a ligand, a Tanimoto coefficient was used to determine if the docked drug was similar to the complexed ligand. A coefficient less than 0.54 represented similar molecules [189], and thus cognate dockings. Evaluation of static RMSD values of protein-drug interactions representing 380 cognate interaction dockings was performed on a case-by-case basis as the chemical numbering of PDB heteroatoms and docked structures often differed, which caused incorrect RMSD calculations. Each RMSD comparison was required to match at least 30% of the docked ligand atoms to the cognate crystal-structure ligand. 320 interactions pass this requirement, of which 221 (69%) showed RMSDs under 2Å. The other 99 (31%) had RMSDs larger than 2Å.

Cytoscape [190] was used to generate the known drug-target interaction map. Networks were fitted to a force-directed layout and manually edited for improved visibility. Drugs and protein targets are nodes in the network, interconnected by interaction edges. The edge lengths were not weighted, and are adjusted for maximum visible understanding.

2.4.5 Applying and evaluating score thresholds

I applied several methods of score thresholding, applying cut-offs for 1) the icm score ranging from 0 to -80 with 16 steps (Δscore of 5), 2) the pmf score ranging from 0 to -250 with 25 steps (Δscore of 10), 3) the drug rank ranging from 1 to 4000 (every rank from 1-15, then 11 steps with increasing intervals until 4000), 4) the protein rank ranging from 1 to 252 (every rank from 1-15, then 7 steps with increasing intervals until 252), and 5) the combined docking score and mean force score cut-offs at each combination of steps

For the consensus score thresholds, all combinations of slopes (from 0 to -40 with step level 0.5) and intercepts (from 0 to -400 with step level 5) were tested. For each line, I calculated the density of the points eliminated in a trapezoidal area delineated by the consensus line, the best icm- score for this protein, the best pmf-score for this protein, the midpoint between the worst icm-score and its mean, and the midpoint between the worst pmf-score and its mean.

Many consensus lines will predict the same number of interactions – the method selects the line that eliminates the densest cloud of poor-scoring points. There are no icm-scores <-123 or pmf-scores <-383 in my dataset; thus these lines are able to cut the dataset and select any X% of drugs. I was careful to note that picking the top 1% of drugs would be uninformative if there were no good scoring binders in the data set, and thus used the midpoint between the worst icm/pmf score and its mean as the minimum score requirement.

2.4.6 Large-scale cross-docking

1,164,492 interactions between 252 proteins and 4621 drugs were docked using ICM. Though there were actually 4854 drugs small molecules, some were excluded being too small or too large for docking (molecular weight under 100 or over 1000 g/mol). Due to the multiple binding pockets per protein and multiple crystal structures per protein, there were a total of 2923 binding pockets. Each interaction was docked 5 times to better cover the docking energy landscape and the best scoring conformation was retained. Overall there were 2923x4621x5 dockings or 68 million docking calculations. The icm and pmf scores of each interaction were gathered into large matrices for further analysis.

2.4.7 Kinase assays

Protein inhibition assays were performed by SignalChem (Richmond, BC, Canada). Kinase assays consisted of ^{33}P-ATP at 5μM, the protein kinase, peptide substrate, assay buffer, and the drug. Blank assays without substrate or drug, and assays without the drug, were used as controls. Staurosporine at 1μM was used as the positive control drug.

Figure 2.1 The computational molecular docking pipeline.

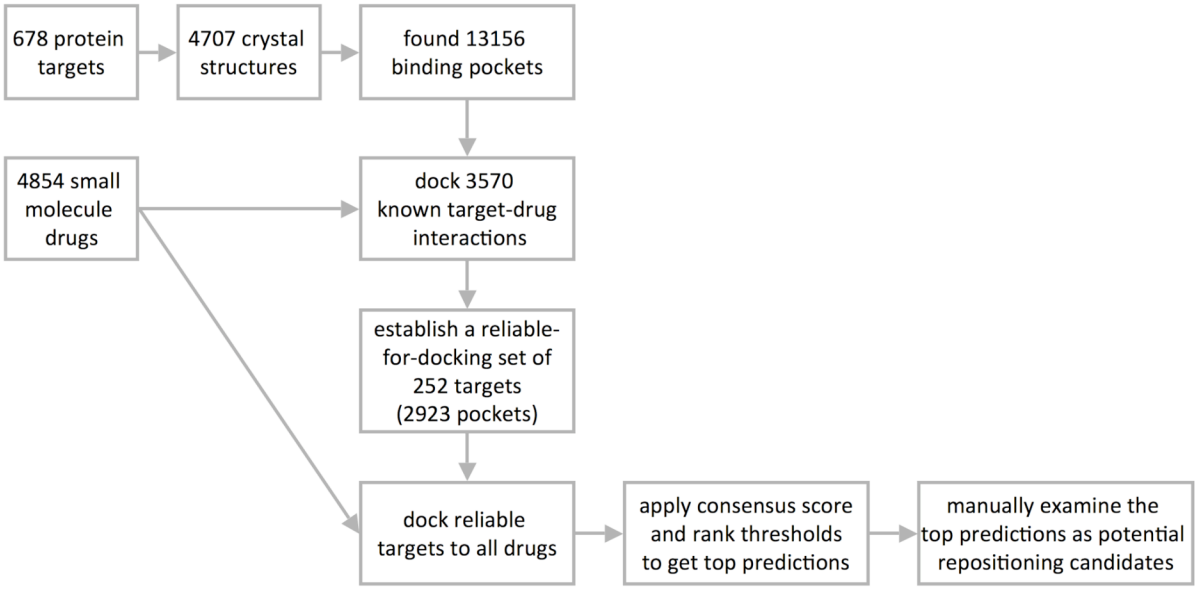


Figure 2.2 Evaluating the known drug-target interaction docking.

1116 (31%) of 3570 known interactions docked with a good score. Two-thirds of the 1116 were ligands docking to non-cognate protein structures, showing that the method could do more than re-dock existing drug-target structures.

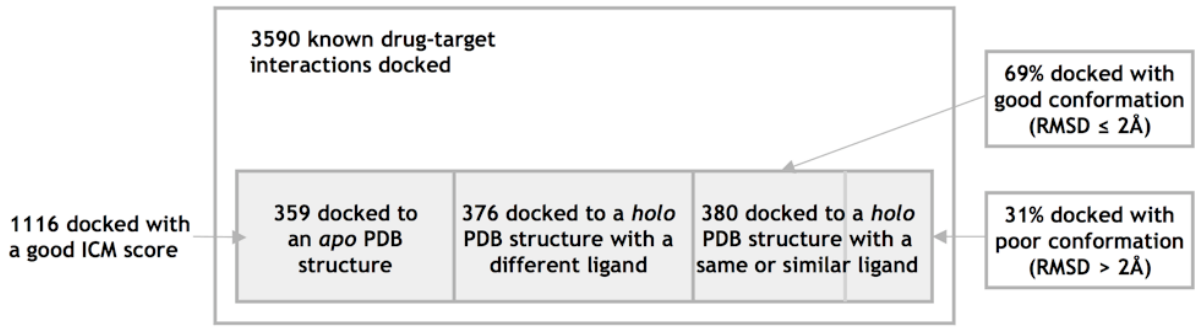


Figure 2.3 Network of known protein-drug interactions.

This network represents the known interactions (grey edges) between proteins (rectangular box nodes), approved drugs (pink circle nodes) and experimental drugs (blue circle nodes). Here are the 252 proteins for which at least one known drug docked well (each protein has at least one red edge) – the 'reliable-for-docking' set. This network shows the high level of interconnection between existing drugs and targets. The proteins at the bottom of the graph are singletons and are not connected to other proteins through shared drugs.

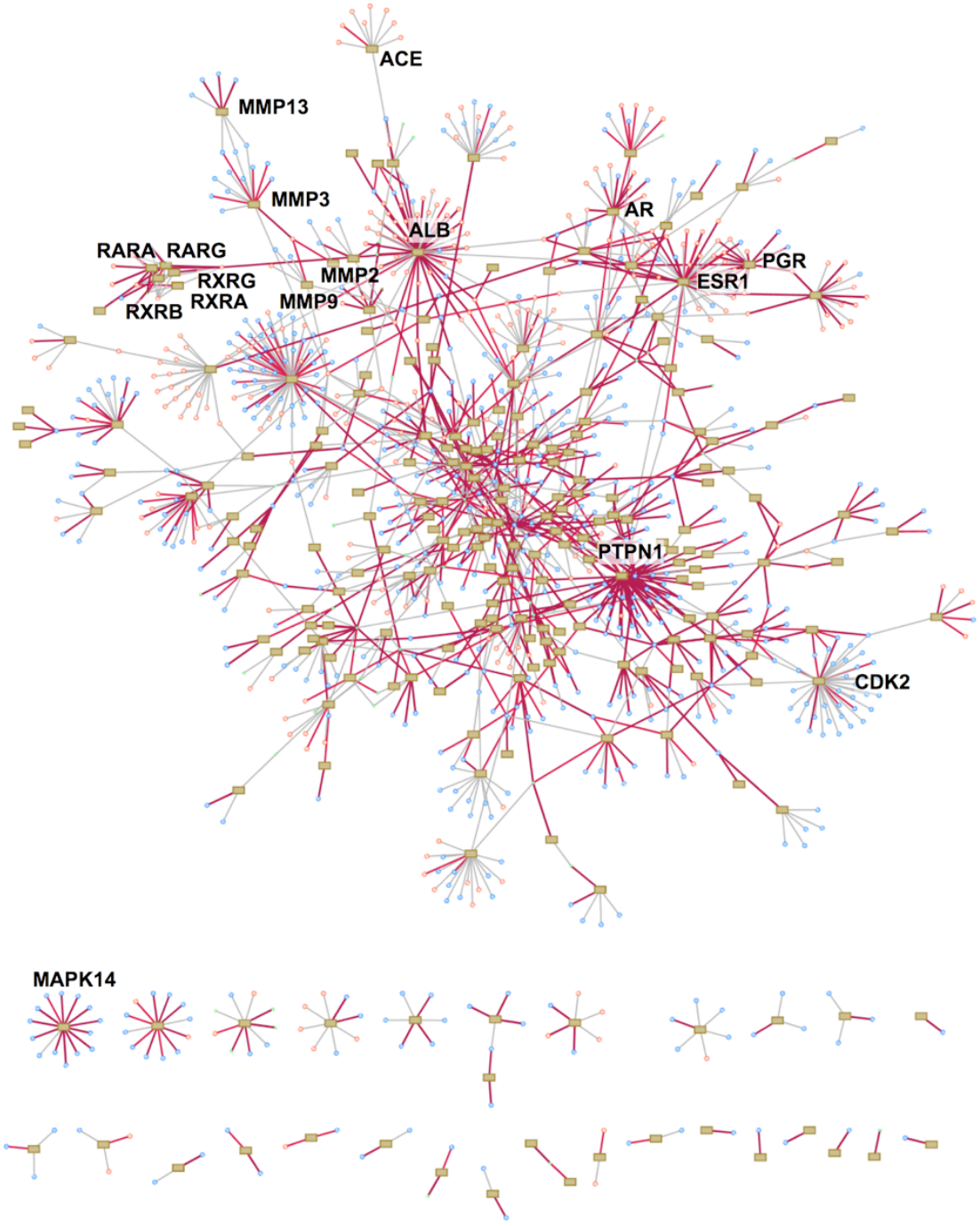


Figure 2.4 Score thresholds assessment

Various combinations of score and rank thresholds were assessed using the positive predictive value (PPV). a) PPVs for thresholds predicting less than 7000 interactions. b) a zoomed in version showing clearer PPV separation for the top 500 predicted interactions.

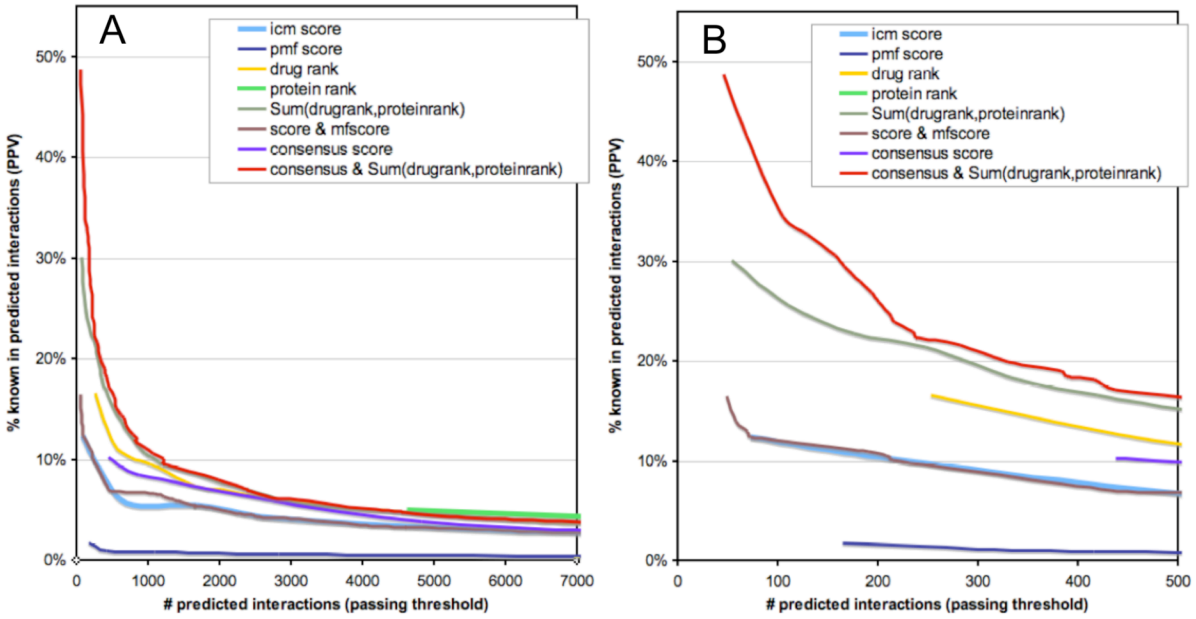


Figure 2.5 The MAPK14 score plot.

Docking icm- and pmf- scores for 4621 drugs to MAPK14. Each point represents a drug. The top 5% of the drugs as determined by the consensus scoring threshold are shown as orange dots. These drugs were also docked to the 252 other drug targets in the database, and circles denote the drugs for which this protein was one of the top 5 targets for the drug. Drugs that are known to bind MAPK14 are shown in red boxes, and it can be seen than most of these red boxes pass both the consensus and protein rank thresholds.

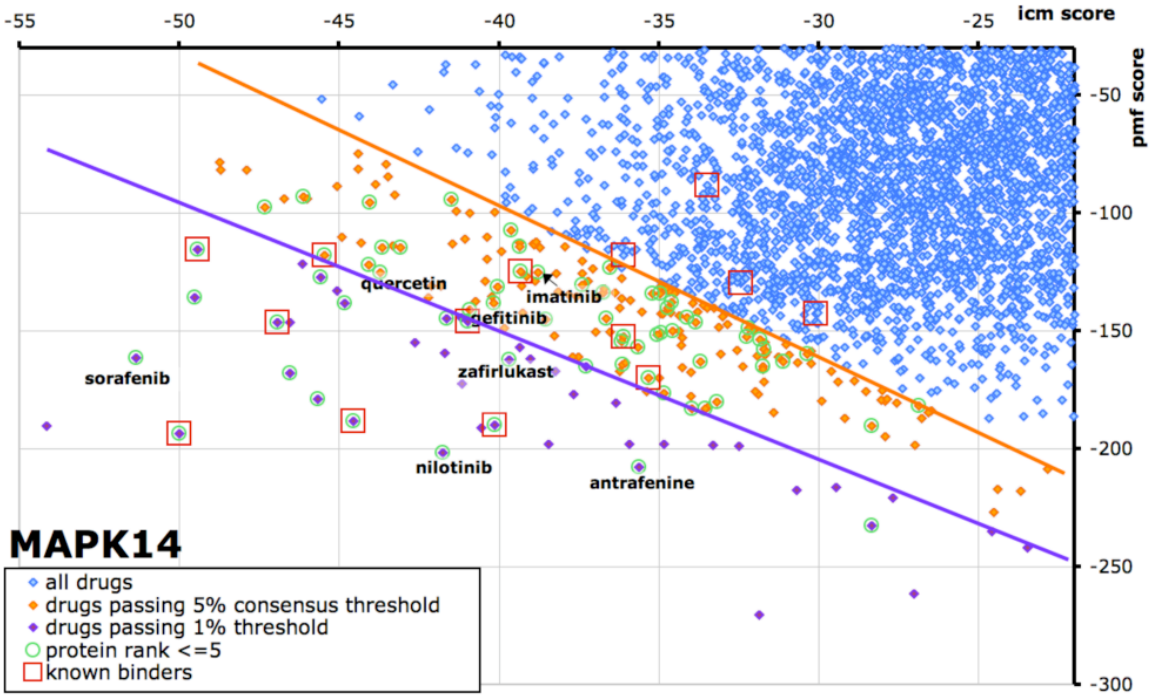


Figure 2.6 Experimental validation of two interactions.

The kinase drug nilotinib (blue) and the asthma drug zafirlukast (yellow) were tested in ATP-competitive enzymatic assays against MAPK14. Results are plotted as percent inhibition of activity versus drug concentration. 1μM staurosporine was used as the positive control. The nilotinib-MAPK14 IC_{50} was calculated to be 40nM.

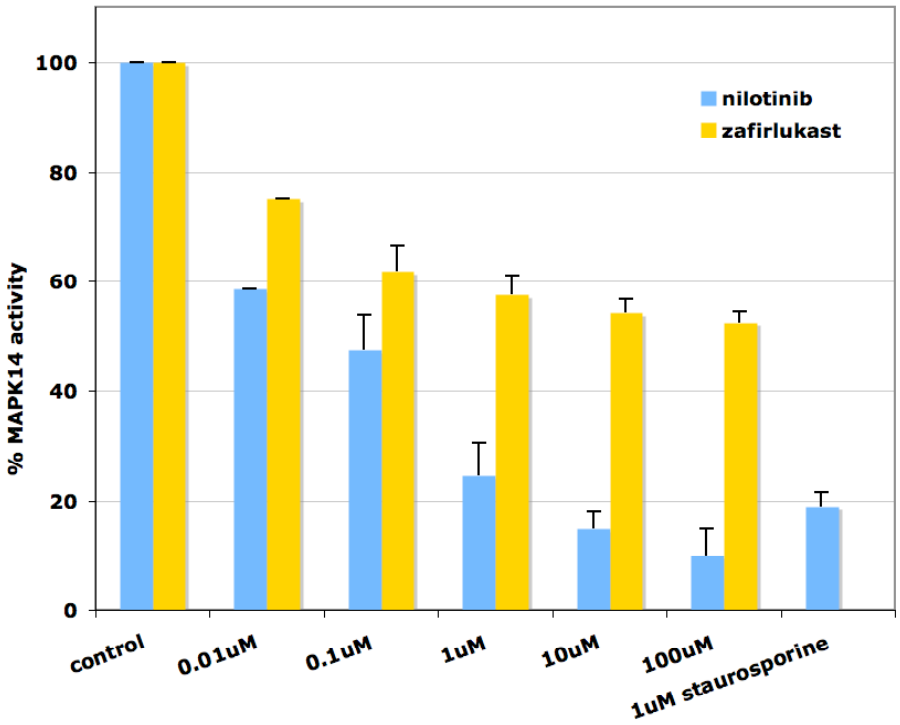


Figure 2.7 The BIM-8 score plot.

Docking icm- and pmf- scores for BIM-8 docked to 252 reliable-for-docking protein targets. Each point represents a protein target. Targets for which BIM-8 passed a consensus threshold are shown as orange dots (top 5%) and brown dots (top 1%). Targets with experimental support are enclosed in red boxes. Targets that have shown no binding activity with BIM-8 in the literature are shown in shades of green. It can be seen that most of the actual targets of BIM-8 pass stringent consensus score thresholds.

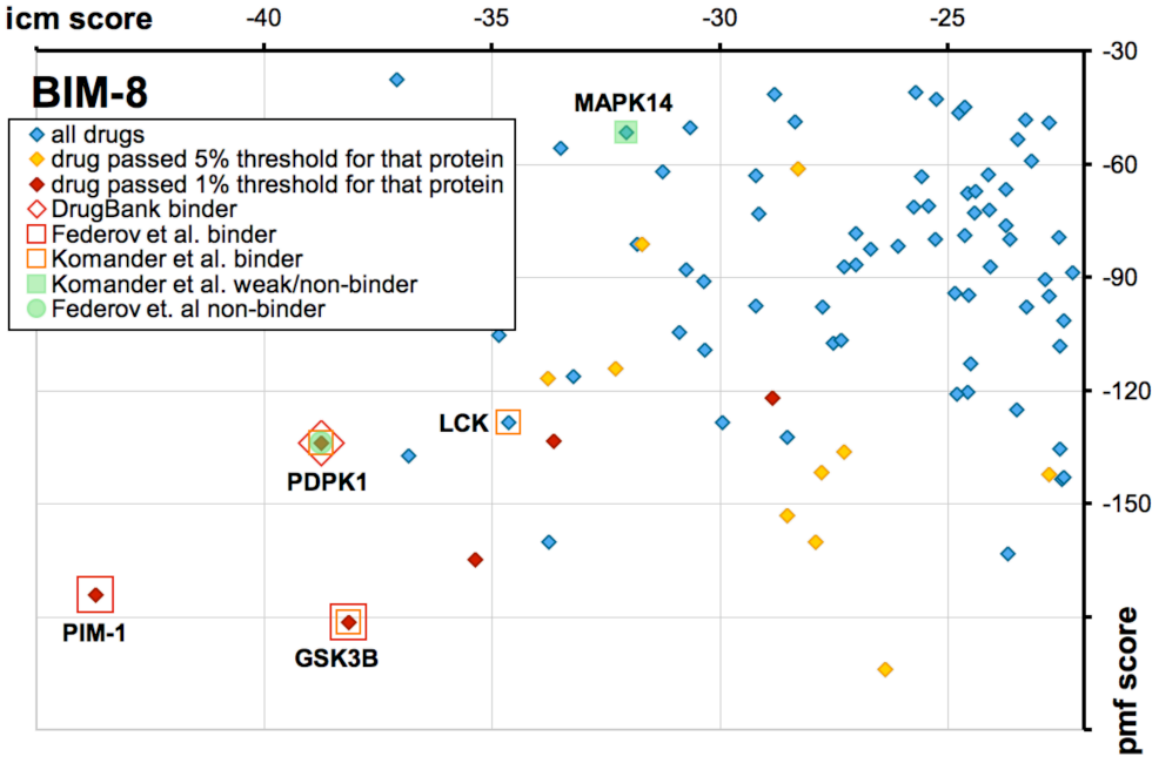


Figure 2.8 Predicted drug-target interaction map.

Quantitative interaction map of drugs docked to protein targets, according to their ICM docking score. Each protein is represented by a row, on which a black cross denotes a known drug docked to the target, a red dot denotes an approved drug docked to the target, and a blue dot denotes an experimental drug docked to the target. Only the top predictions for established drug targets (at least one known approved drug) that docked with a score passing the consensus threshold and had a protein-rank ≤5 are shown.

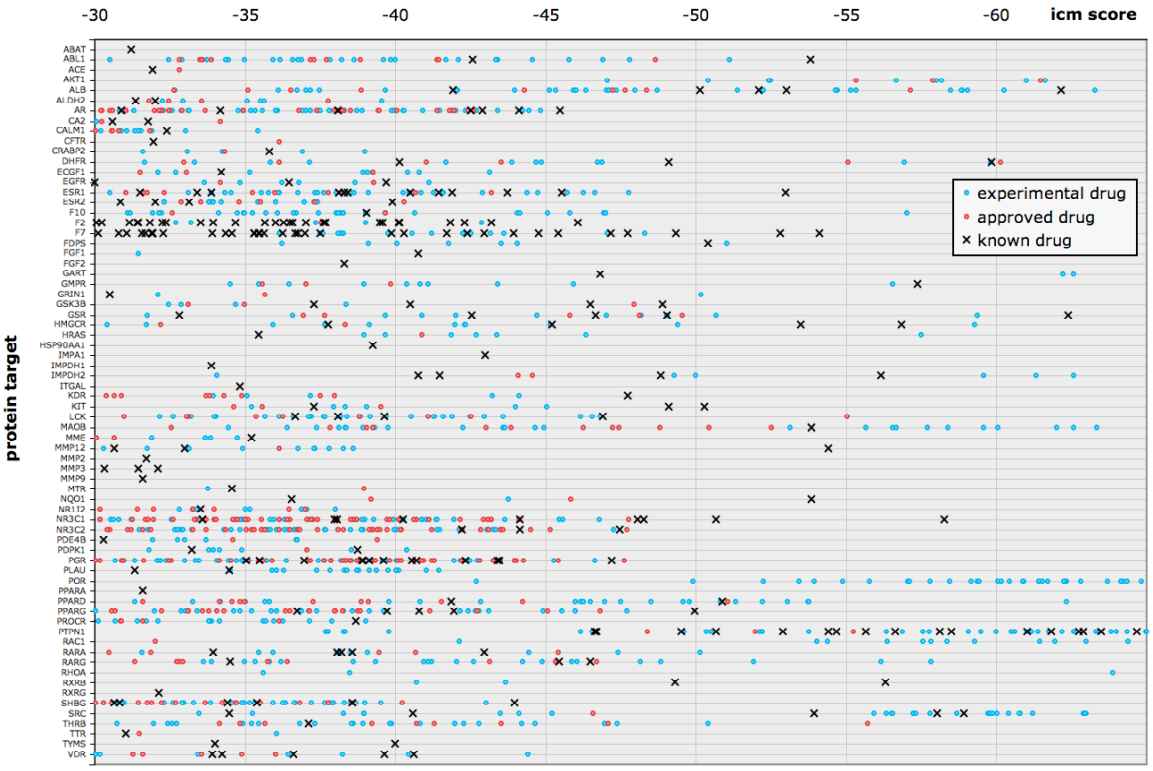


Table 2.1 A comparison of various threshold methods.

Testing the ability of various threshold methods to predict a high percentage of known interactions (PPV) and enrich the predicted interaction set for known interactions. The sum rank is the sum of the protein and drug ranks for that interaction. Thresholds are listed by increasing enrichment.

Threshold	# predicted interactions	# known in predicted interactions	# proteins in interactions	% known in predicted set (PPV)	enrichment factor vs random
random	1,164,492	1116	252	0.1%	1
icm-score of -30	104,625	1116	252	1.1%	11
pmf-score of -300	150	3	20	2.0%	21
protein-rank of 1	4621	234	206	5.1%	53
consensus score 0.05%	437	45	238	10.3%	107
icm-score of -100	72	9	17	12.5%	130
drug-rank of 1	252	42	252	16.7%	174
icm-score -100 & pmf score -140	48	8	13	16.6%	174
drug rank 1 & protein rank 1	53	16	53	30.2%	315
consensus score 0.05% & sum rank ≤4	45	22	39	48.8%	510

Table 2.2 A comparison of various threshold methods.

A comparison of various threshold methods based on their ability to predict a high percentage of known interactions (PPV) and enrich the predicted interaction set for known interactions compared to other methods.

Threshold	# predicted interactions	# known in predicted interactions	# proteins in interactions	% known in predicted set (PPV)	enrichment factor vs random
use icm- score of worst scoring known binder	62337	1117	252	1.8%	20
use icm- & pmf- scores of worst scoring known binder	28840	716	252	2.5%	27
use icm- score of best scoring known binder	16412	253	252	1.5%	17
use icm- & pmf- scores of best scoring known binder	7859	253	252	3.2%	35

Table 2.3 The ability of thresholds to enrich for known MAPK14 inhibitors.

	All docked drugs	**Known drugs ligands**	**Enrichment factor versus random**
# docked to MAPK14	4621	14	1
# passing icm score ≤-30	970	14	5
# passing 5% consensus score	225	10	15
# passing 5% consensus & protein rank ≤5	67	10	49
# passing 1% consensus score	45	6	44
# passing 1% consensus & protein rank ≤5	18	6	110

Table 2.4 The ability of various thresholds to enrich for BIM-8 targets.

	All docked drugs	**Known drugs ligands**	**Enrichment factor versus random**
# proteins BIM-8 was docked to	252	4	1.0
# passing default score ≤-30	24	4	10.5
# passing 5% consensus score	20	4	12.6
# passing 1% consensus score	6	3	31.5
# passing 5% consensus & protein rank ≤5	3	3	63
# passing 1% consensus & protein rank ≤5	3	3	63

Table 2.5 Top predicted hits that have literature support.

Protein	Drug	icm score	pmf score	drug rank	prot rank	Notes
AIFM1	DB02332	-79	-231	1	1	Flavin is a cofactor. [191]
ALB	DB03756	-66	-163	1	2	Dosahexanoic acid (DHA) can form complex with albumin and confers neuroprotective effects in rats. [180]
ALB	DB06689	-51	-130	84	3	Ethanolamine oleate promptly binds with albumin in the blood. [17]
AKT1	DB03265	-81	-95	2	1	Crystal structure of inositol 1,3,4,5-tetrakisphosphate bound to AKT. [16]
BTK	DB03344	-69	-99	1	3	[192] shows that inositol 1,3,4,5-tetrakisphosphate binds to BTK. This compound is very similar: inositol 1,3,4,5,6 tetrakisphosphate.
CYB5R3	DB02332	-71	-258	2	2	Flavin is a cofactor. [193]
ESR1	DB05414	-47	-197	3	1	ERA-923 is a selective estrogen receptor modulator. [172]
ESR1	DB01645	-42	-109	16	1	Genistein is a selective estrogen receptor modulator. [173]
GART	DB02223	-63	-126	1	5	LY-231514 tetra-glu a known thymidylate synthase inhibitor. LY-231514 is pemetrexed, a GART and thymidylate sythase inhibitor. inhibitor. [194]
GART	DB02794	-62	-147	2	4	Crystal structure of compound bound to E.coli GART. [195]
GSR	DB02332	-57	-211			Flavin is a cofactor. [191]
KDR	DB04879	-49	-152	1	1	Vatalanib is a pan VEGFR inhibitor. IC_{50} 37nM. [177]
KIT	DB04868	-44	-240	4	2	Nilotinib binds to KIT. [178]
MAPK10	DB00317	-39	-183	72	3	Gefitinib binds MAPK10 weakly: Kd=2-3μM. [42]
MAPK14	DB00398	-51	-161	2	2	Sorafenib IC_{50} 0.057μM. [163]
MMP2	DB02255	-37	-84	1	6	Illomastat is a broad-spectrum MMP inhibitor. Ki 0.5nM (Chemicon International Inc, Temecula, CA)
MMP8	DB02255	-44	-67	2	1	Illomastat is a broad-spectrum MMP inhibitor. Ki 0.1nM (Chemicon International Inc, Temecula, CA)

Protein	Drug	icm score	pmf score	drug rank	prot rank	Notes
NR3C2	DB01395	-48	-150	1	1	Drospirenone, a progestogen with antimineralocorticoid properties. [196]
PPARD	DB03756	-62	-144	1	4	DHA can activate PPARD. [197]
PPARG	DB06536	-47	-130	9	1	Tesaglitazir is a dual PPARA/PPARG agonist. [198]
RAC1	DB03532	-120	-145	1	1	RAC1 is a GTPase [191], and this compound is a standard GTP analog.
RARG	DB02466	-58	-216	1	1	BMS181156 binds RARG with Kd 0.6nM. [199]
RARG	DB02258	-56	-220	2	1	SR11254 is a RARG-selective ligand. [200]
RARA	DB05076	-45	-131	6	2	4-HPR is a highly selective activator of retinoid receptors. [201]
RARG	DB05076	-46	-134	6	1	4-HPR is a highly selective activator of retinoid receptors. [201]
RARG	DB02741	-52	-217	3	1	CD564 binds RARG with Kd 3nM. [199]
RARG	DB03466	-46	-208	11	1	BMS184394. [199]
RXRA	DB03756	-54	-137	1	8	DHA. [202]
RXRA	DB04557	-53	-156	2	5	Arachidonic acid. [202]
VDR	DB04891	-49	-204	1	1	Becocalcidiol, a vitamin D analog. [174]
VDR	DB04295	-44	-297	4	1	ED-71, a vitamin D analog. [175]

3 Identifying Novel EGFR Inhibitors by Computational Drug Repositioning Analysis

3.1 Introduction

The tyrosine kinase EGFR is a member of the ErbB family of cell surface receptors. Upon binding of the epidermal growth factor (EGF), EGFR dimerization and activation is induced. This leads to the initiation of many signaling pathways including the MAPK (RAS/RAF/MEK/ERK) and the PI3K pathway (PI3K/AKT/MTOR), among others [203]. As these pathways play key roles in cell survival and proliferation; EGFR has been deeply studied as a drug target in cancer research. Antagonists of EGFR are approved for the treatment of non-small-cell lung cancer, head and neck cancer, colorectal cancer, and pancreatic cancer [204]. The antagonists are also under evaluation for treatment of breast cancer, ovarian cancer, and renal cell carcinoma [204].

EGFR is also an attractive target for molecular docking studies, with crystal structures solved in a variety of conformations as well as numerous known ligands. To date, several studies have searched for novel EGFR inhibitors using docking approaches. Cavasotto *et al*. docked 315,102 compounds of the ChemBridge Express Library against an EGFR crystal structure and found seven compounds with 40-50% inhibition of kinase activity at 10μM [158]. Choowongkomon *et al*. docked the NCI Diversity set of 1990 compounds against one EGFR crystal structure and found 8 potential interactions [205], although these interactions were not validated in binding assays. Recently Li *et al*. used support vector machines to create a scoring function for EGFR, and docked an in-house library of 1125 compounds to one EGFR crystal structure [206]. They found 3 compounds with direct binding IC_{50}'s of 2, 10, and 56 μM. La Motta *et al*. performed QSAR and docking studies to EGFR with the Maybridge database and found that docking studies performed best, identifying two low micromolar inhibitors of EGFR [207]. Taken together, these studies suggested that docking could be an effective approach for identifying novel EGFR inhibitors.

I chose to search for an existing drug that could be repositioned as an EGFR inhibitor, for potential use as a cancer treatment. Here I used molecular docking to 23 EGFR crystal structures followed by stringent false-positive filtering to predict such repositioning candidates. My computational method was able to predict known EGFR binders from a drug database with high enrichment. I then virtually screened the DrugBank chemical library against EGFR and experimentally tested one of the top hits. Though the drug appeared to inhibit both EGFR pathway signaling and EGFR-overexpressing cell line proliferation, it ultimately did not inhibit EGFR in ATP-competitive kinase binding assays. I thus concluded that this result was a false positive docking prediction, and summarized the limitations of docking methods that would be important to consider in future VS analyses.

3.2 Results

3.2.1 Molecular docking to EGFR structures

Three-dimensional crystal structures of EGFR collected from the Protein Data Bank [76] are listed in Table 3.1. There were a total of 23 structures: 11 with wild type kinase domains, 8 with the G719S or L858R mutation, and 4 with only the extracellular domain. 65 binding pockets were predicted in these structures, including the known ATP-binding pockets as well as potentially novel pockets in the kinase and extracellular domains (Figure 3.1). I docked 5908 small molecule drugs from three drug databases to the predicted EGFR binding pockets using the ICM software package [87]. The small molecules included 806 FDA approved drugs, 768 experimental drugs, and 39 nutraceutical drugs from DrugBank [144]; 3102 drugs approved in USA and Japan from KEGG DRUG [145]; and 1193 oral marketed drugs collected by a 2004 study [208]. As described in Chapter 2, each docked protein-drug interaction had two independent scores describing its binding potential: an energy-based score (icm-score) and a knowledge-based potential of mean force score (pmf-score).

3.2.2 Filtering the docking results

Figure 3.2 plots the two scores for each compound docked to EGFR and shows that few compounds have both good icm- and pmf-scores, located in the bottom-left of the plot. I applied the consensus scoring method developed in Chapter 2 that considers each docked interaction as a pair (icm-score, pmf-score), and removes the densest group of poor scoring

pairs on the plot. I selected the threshold line that minimally allowed the known EGFR-drug canertinib to pass. This resulted in 50 candidate drugs including all 7 known EGFR binding compounds (Table 3.2).

I then docked these top 50 drugs against a reliable-for-docking set of protein targets consisting of 2231 human protein crystal structures (134 unique drug targets, an earlier version of the dataset described in Chapter 2). Targets were ranked by their icm-score to each drug, thereby creating a 'protein rank,' which measured the selectivity of the ligand to the protein. I filtered out compounds for which EGFR had a protein rank of greater than 3 (EGFR was not one of their top three predicted targets), resulting in just 20 drugs. Many high scoring drugs were eliminated at this step. These predicted EGFR-binding drugs are listed in Table 3.3.

There are other, more standard ways of choosing top docking hits (Table 3.2 first two rows and Figure 3.2 blue lines). Using the default software-specified score threshold, an icm-score of -32, resulted in 528 hits. More lenient thresholds can also be used; for example, the EGFR study by Cavasotto *et al*. used a score threshold of -28 [158]. Using the worst known-drug docking score as the cut-off (gefitinib with an icm-score of -36) resulted in 150 hits. In comparison, my computational method produced a much more compact list of 20 predicted binders. However, in order to be useful, this method was required to enrich the predicted set for known binders.

3.2.3 Analysis of known EGFR inhibitors

A standard measure of docking success for a target is to evaluate the docking of its known inhibitors as a positive control. In our my database there were four known EGFR drugs (gefitinib, erlotinib, lapatinib, canertinib), two metabolites (ADP, ATP), and a broad-spectrum kinase inhibitor (staurosporine). They all had high scores passing the consensus score threshold (Figure 3.2). The four known drugs also passed the protein rank criteria; thus, EGFR was one of the top three proteins for these drugs when docked to 134 drug targets. This property did not hold for ADP, ATP, and staurosporine with EGFR protein ranks of 4,

6, and 8, respectively. However, these compounds are known bind to all kinases (or >90% of kinases in the case of staurosporine [42]).

The drugs' predicted binding conformations were also compared to their respective poses in EGFR crystal structure complexes, a couple of which are shown in Figure 3.2c. I considered a heavy atom RMSD under 2Å to be a measure of successful docking, and under 1Å to be a very rigorous measure of docking accuracy [97]. For example, the docked ADP was compared to the ATP-analog AMP-PNP (PDB code 2ITX) where RMSD is 0.8Å. Although staurosporine obtained the lowest score out of the known EGFR binders, its RMSD to the analog AFN941 (PDB code 2ITQ) was still successful at 1.6Å. The known drugs all exhibited RMSDs under 2Å and variation occurred mostly in the solvent-exposed regions of the bound ligand that did not directly interact with the protein. The core positions of the bound and docked ligands are very similar, with RMSDs under 1Å. In addition, hydrogen bonds formed in the crystal structure complexes were retained in the docked binding conformations. In short, the computational method successfully reproduced all the known EGFR ligands both in terms of score and conformation. I therefore applied my method to predict potential drug repositioning candidates for EGFR.

3.2.4 Analysis of top predicted drug repositioning candidates of EGFR

The 20 predicted EGFR binding compounds passing all score and protein rank thresholds are shown in Table 3.2. Aside from the 4 known binders, several kinase targeting compounds were also predicted. Of note is compound 2, a CDK2 inhibitor [209] that is likely to inhibit EGFR as it belongs to the same 4-anilinoquinazoline class of compounds as gefitinib, erlotinib, and canertinib. Atorvastatin is predicted to bind to EGFR at a site on its extracellular domain, and a previous study has shown that this drug inhibits the phosphorylation of EGFR and ERK in mice [210].

An additional result of performing a protein-rank analysis is the prediction of a protein-drug network centered on EGFR (Figure 3.3), showing drug targets that are connected to EGFR through one drug. There are many kinases in the network (CDK2, MET, PDPK1, LCK, HCK, MAPK10, MAPK14, ABL1, and KIT), which is in accordance with ATP-binding sites

being similar across kinases and thus being expected to bind drugs in common with EGFR. In particular, there are many compounds shared between HCK and LCK, which is also likely considering 26 of 28 binding site residues of these two protein targets are identical. In short, this network shows us other proteins that could be targeted in combination with EGFR using a single drug.

I selected one prediction for experimental validation, requiring it to be an approved drug and have a very good (icm-score, pmf-score) pair. Since compounds with higher pmf-scores have more statistically likely interactions with protein atoms, I selected the anti-HIV drug tenofovir disoproxil fumarate (TDF) (Figure 3.4a) known to inhibit reverse transcriptase. TDF passed the consensus thresholds and also had a very high pmf-score (Figure 3.2a). This compound docked well to both the wild-type and mutant structures of EGFR. Despite being chemically similar to ADP, TDF has a different predicted binding conformation (Figure 3.4c). Like ATP, TDF is predicted to establish hydrogen bonds with M793; however, it is also predicted to form two hydrogen bonds to K745, a residue that is known to be critical for EGFR activity [211]. These extra hydrogen bonds anchor TDF in the ATP-binding site from both ends, interacting with additional residues at the ATP site. TDF appeared to be a good candidate for EGFR inhibition from both scoring and posing aspects, and I followed with *in vitro* experiments on TDF for validation.

In the protein-drug network, TDF also docked well to two other non-kinase targets. The first target UDP-galactose-4-epimerase (GALE) is one of three human galactose-metabolizing enzymes. Individuals born with defective GALE develop galactosemia and must be treated with a low galactose diet [212]. Interestingly, TDF therapy in AIDS individuals has been increasingly recognized as a cause of acquired Fanconi's syndrome (FS), which in turn can be caused by cystinosis, Wilson's disease, tyrosenemia, as well as galactosemia [213, 214]. Though FS is an uncommon adverse effect of TDF, the GALE protein may play a role in these individuals. The other target is the repair enzyme protein-L-isoaspartate (D-aspartate) O-methyltransferase (PIMT). In mice, PIMT-deficiency results in fatal epilepsy [215], but it does not have an elucidated role in human diseases to date. TDF is not known to cause epilepsy in AIDS patients. TDF is not predicted to interact with other kinase proteins and

may thus have a very different off-target profile from current EGFR drugs. The predicted specificity of TDF for EGFR may lie in its binding conformation (Figure 3.3c), interacting with three residues spanning the ATP binding site. In contrast, the known drug erlotinib only had one hydrogen bond to EGFR (Figure 3.2b).

3.2.5 TDF inhibits cell proliferation of breast cancer cell lines

Since TDF is predicted to target the EGFR ATP-binding site, it was expected to inhibit cells in a manner similar to the known EGFR drug gefitinib (marketed as Iressa®; AstraZeneca). I thus investigated the effect of TDF on three gefitinib-sensitive cell lines. The first two are known to overexpress EGFR: the basal-like breast cancer (BLBC) cell line SUM149, known to overexpress wild-type EGFR [216] and the human epidermoid carcinoma A431 cell line [217]. The third was metastatic breast cancer cell line BT474M1 which expresses wild-type EGFR but is also known to overexpress HER2 [218]. After 72 hours of treatment with TDF, cell growth was significantly inhibited up to 35% at 1μM and 57% at 10μM in SUM149 (Figure 3.5a). The inhibition was dose-dependent with an estimated cellular IC_{50} of 8.1μM. It was not as potent as gefitinib (cellular IC_{50} ~1μM, which is in accordance with previous studies [216]. A431 and BT474M1 cells also showed similar inhibition % at 10μM (note that the A431 cells were not tested at 1μM).

I tested TDF from pharmacy acquired pills, in the form of Truvada (TDF + emtricitabine) and found that Truvada inhibited SUM149 cells with an IC_{50} of 3.5μM, but did not show any inhibition on the gefitinib-insensitive cell lines (MDA-MB231) [219] (Figure 3.5b). The IC_{50} values were estimated to be 46μM and 71μM, respectively. As a positive control, I determined that gefitinib was able to inhibit SUM149 cells at 1μM but only weakly inhibited MDA-MB231 and HCC1937 cells. Overall, these experiments supported the hypothesis that TDF was inhibiting the EGFR protein in a manner similar to the known EGFR-targeting drug gefitinib.

I also tested TDF on the melanoma cell line LCC6 (Figure 3.5c) which expresses wild type EGFR [220]. In this cell line, it is thought that HER2 overexpression is the main cancer driving mechanism. However, the EGFR-HER2 interaction causes prolonged HER2

signaling and in accord, LCC6 cells are known to be sensitive to the EGFR-drug gefitinib [220]. Though I did not observe strong inhibition of LCC6 by gefitinib, studies performd by Warburton *et al*. showed that the concentration of fetal bovine serum (FBS) was important in such an assay, since gefitinib did not inhibit LCC6 cells at 10% FBS. However, 1μM gefitinib caused a 66% inhibition of cell proliferation compared to untreated controls at reduced serum conditions (1% FBS), suggesting that an excess of growth factors may override the low drug concentration. The 5% FBS present in my assay may have been too high to observe gefitinib inhibition; however, TDF still showed a strong dose-dependent inhibition of LCC6 proliferation.

3.2.6 TDF inhibits EGFR pathway signaling in an EGF-driven manner

I investigated the effects of TDF compared to gefitinib on the signaling of EGFR and ERK, a key protein in the EGFR signaling pathway. TDF treatment for 24 hours significantly decreased phospho-ERK 1/2 protein levels while total-ERK protein levels remained constant (Figure 3.6a). The control drug gefitinib also decreased ERK signaling. However, ERK signaling can also be decreased when receptor tyrosine kinases other than EGFR are inhibited. Thus in order to show that the TDF inhibition was EGF-dependent, I stimulated serum-starved SUM149 cells with EGF to see whether signaling was still inhibited (Figure 3.6b). As expected, following EGF stimulation, phosphorylation of ERK increased and gefitinib was able to suppress this (Figure 3.6b left). TDF also showed strong EGF-dependence evidenced by the dramatic inhibition of ERK phosphorylation following EGF stimulation (Figure 3.6b right). Thus it appears that the anti-proliferative activity of TDF is a result of directly inhibiting EGFR signaling in an EGF-driven manner.

3.2.7 TDF does not inhibit EGFR in direct binding assays

After completing the cell line and signaling experiments, an *in vitro* EGFR assay became available at SignalChem (Richmond, Canada). Though preliminary results seemed promising (40% inhibition at 10nM drug concentration), subsequent replicates did not show any inhibition or dose-dependent interaction for TDF (Figure 3.7a). I speculated whether the peptide used in their assay may be interfering with TDF binding. Therefore, I also conducted a kinase assays from Invitrogen and Caliper, which used different peptides and fluorescent

detection methods; however TDF did not inhibit EGFR any more than baseline variation in either assay (Figure 3.7c). In short, TDF was not able to inhibit EGFR in three different biochemical binding assays and I concluded that its effects on MAPK signaling inhibition may have been through other protein targets.

One important lesson resulting from these experiments was the importance of the ATP concentration in the kinase assays (Figure 3.7b). SignalChem uses standard settings of 50μM ATP with staurosporine as the control drug. However, my control drug, the known nanomolar inhibitor gefitinib, did not show any inhibition in such conditions. At the lower ATP concentration of 5μM, gefitinib was able to display a clear dose-dependent response.

3.3 Discussion

In this study, I have developed a docking pipeline for EGFR using an ensemble of 23 existing EGFR crystal structures. With a set of seven known binders (including 4 approved drugs), the scoring and ranking criteria were able to obtain enrichment factors over 100. The most stringent criteria had a PPV of 4/20=20%. Extrapolating, there would be a 20% chance for the chosen drug TDF to be a true EGFR inhibitor.

That TDF was not observed to inhibit EGFR directly was disappointing given the previous experimental data showing that TDF could inhibit ERK phosphorylation and was dependent upon the presence of EGF. There were several possible explanations for the contradictory results. One possibility could be some type of chemical aggregate forming during assay conditions that prevents it from binding to EGFR. However, this is very unlikely as aggregates would not be able to penetrate the cells and inhibit the EGFR pathway signaling. TDF may be inhibiting another protein in the cell that can affect ERK phosphorylation. For example, Protein Kinase C is known to inhibit ERK phosphorylation independent of Ras, an important protein along the EGFR pathway [221]. However, this is also unlikely as TDF inhibition was observed to be EGF-dependent. If TDF was inhibiting through PKC, I would not expect to see such strong dependence on EGF.

TDF could also be inhibiting the proteins PIMT or GALE. PIMT knockdown has been previously associated with hyperactivation of EGF-stimulated MEK and ERK signaling in mammalian cells [222]. Since TDF treatment inhibited ERK signaling in breast cancer cells, I do not believe it inhibited PIMT in these cells. Though there is indirect support for the TDF-GALE interaction through galactosemia, I could not find any links between GALE and ERK signaling in the literature, and thus do not believe GALE is involved in the observed inhibition of ERK phosphorylation.

I considered whether TDF was inhibiting one of the other ERBB family proteins (other than EGFR, which is also known as HER1/ERBB1, there are HER2/ERBB2, HER3/ERBB3, and HER4/ERBB4). EGF is an agonist specific to EGFR; however, EGFR is known to form a homo- or hetero- dimer with one of the other family members before continuing the signaling cascade [204]. A sequence comparison of the family members (Figure 3.8) shows that they share many identical residues (shaded in dark green), especially within 3.5A of the ATP binding site (boxed in red). Though there were no solved PDB structures of these other structures at the time, it is likely that TDF would be also able to inhibit one or more of them. It is interesting to note that TDF could inhibit LCC6 cells at 1μM, whereas gefitinib could not. This may suggest some interaction between TDF and HER2 that allowed it to more strongly inhibit the LCC6 cells.

TDF could also be inhibiting another protein along the EGFR pathway, downstream of EGFR but upstream of ERK, or even perhaps ERK itself. Such proteins include Ras, Raf (B-Raf, C-Raf), MEKs (MEK1/2), ERKs (ERK1/2) [223]. These proteins were not in my database as their crystal structures had not yet been solved; thus, I was not able to include them in the inverse docking study. I believe that this low representation of all proteins in the cell is currently the biggest limitation of my method – as well as all other drug-target interaction prediction methods. If the true target of TDF were in the dataset, the TDF-EGFR interaction would have a poorer protein-rank and thus fail the protein-rank thresholds. As more crystal structures of proteins are solved, the accuracy of the protein rank filter as well as the protein-drug network can be improved. Of course, improvements in docking scoring mechanisms will also be critical for improving the utility of the method (i.e. improving the

score of staurosporine). As scoring methods improve in accuracy, I would also expect the docking score of TDF to be lower and thus fail the consensus score threshold.

Finally, it could be that all three direct binding assays were not able to detect the inhibition of EGFR by TDF. It is possible to use other assays such as washing cell lysates over column-bound TDF; however, this method does not generate IC50 values that are necessary for any potential therapeutic. If the opportunity arises, it would also be informative to test the TDF-EGFR interaction using BIACORE assays, which do not rely on fluorescence readings but instead detect changes in protein mass upon ligand binding[102].

The method performed well for many of the known EGFR binders. The goal was to filter out the majority of false positive predictions that are predicted by docking. The large enrichment values in this study suggest that many false positive predictions have been eliminated, and the PPV values suggest that 20% of the of the remaining 19 predicted hits may be real EGFR binding compounds. I was not able to validate TDF, but noted that it did not fall in an optimal section of the score plot (i.e. bottom right). It was also possible that there were few strong inhibitors in the library and that my stringent thresholds eliminated all the true positives. Determining the cause will require experimental testing of the other top predictions. However, TDF was one of the few drugs that could be purchased due to its approved status, in contrast to experimental drugs synthesized in specific labs. Overall, the results of this study indicate that there may be few strong EGFR inhibitors within the existing approved drugs aside from the existing inhibitors and that I should apply the method to screen larger databases of chemicals to find novel EGFR inhibitors. Drug repositioning currently may not be a viable strategy for targeting EGFR-associated diseases.

3.4 Methods

3.4.1 Known drug database collection

Chemical structures in SDF format were obtained from DrugBank, consisting of 806 FDA approved drugs, 768 experimental drugs, and 39 nutraceutical drugs [144]; 3102 drugs approved in USA and Japan from KEGG DRUG [145]; and 1193 oral marketed drugs

collected by a 2004 study [208]. The DrugBank compounds all had at least one known human protein target.

3.4.2 EGFR crystal structures collation

EGFR crystal structures were obtained from the Protein Data Bank (PDB) [224]. I prepared protein structures for docking using Molsoft's ICM software version 3.6-1c [87], removing water molecules, solvent ions, and other ligands from the structures. I added hydrogen atoms to the structures then optimized in their positions. To predict potential ligand-binding pockets in the proteins, I used the PocketFinder [78] method in ICM, which calculates a transformation of the van der Waals energy for an aliphatic carbon probe on a grid map. The receptor is defined as the box 3.5Å surrounding the pocket, and the three largest pockets were added to the EGFR binding site collection. In total, there were 65 pockets.

3.4.3 Protein drug target crystal structures collection

I obtained human protein drug targets from DrugBank. PDB structures with at least 95% sequence identity for proteins at least 20 amino acids in size were obtained. Structures were required to be X-ray crystal structures with a minimum resolution of 2.5Å. Multiple chains were grouped into a set of non-redundant sequences, based on PDB's chain redundancy analysis at the 95% sequence identity level [225]. As with the EGFR structures, I added hydrogen atoms, predicted pockets, and defined receptor areas.

3.4.4 Molecular docking

Drugs were docked to target receptors using the ICM virtual library screening (VLS) module. This method performs rigid-receptor flexible-ligand docking using a two-step Monte Carlo minimization method and energy scoring function to sample ligand conformations and select the best docking hits. MMFF partial charges and ECEPP/3 force-field parameters were used. To ensure a sufficient coverage of the docking energy landscape, I docked every drug-target interaction 10 times. High-throughput docking was performed on a Linux cluster with 175 licenses of ICM. As a given protein may have several structures, each of which with more than one pocket, I docked all pockets to a drug, and the best scoring interaction is selected to be the representative protein-drug score.

3.4.5 Consensus score threshold

I set the minimum icm- and pmf- score threshold to be (-30,-30). On a 2D score plot, straight lines of negative slope and intercept were used to separate the predicted hits into passing-threshold and failing-threshold groups. As seen in Figure 3.2, my best predictions passing the threshold were in the bottom left corner of the plot. The failing predictions fell in the trapezoid formed by the line, icm-score=-30, pmf-score=-30 and the minimum and maximum icm- and pmf- scores. For each line, I calculated the density of failing-threshold hits in the trapezoid. Of the lines that just included canertinib (which had the worst score pair of the known drugs) in the passing group, I chose the line that failed the densest cluster of points.

3.4.6 Inverse docking

I docked the 50 compounds that passed the consensus scoring threshold to a database of 2231 binding pockets for 134 unique human protein drug targets annotated by DrugBank. This database was gathered as per 3.4.3 and is an earlier version of the pocket database described in Chapter 2. It consists of target proteins for which at least one known drug could be docked with a good score. The database has been described in my previous work [226], and represents a selection of targets that are more reliable for docking. Since then, I have updated it with the latest DrugBank and PDB information. Inverse docking was carried out as the docking in 3.4.4, though each drug-target interaction was docked 5 times instead of 10. Since I could not perform a icm-, pmf- score threshold for each protein (each protein was only docked to 50 drugs, not allowing enough points for a reliable autonomous threshold), I applied a simple -32 score threshold.

3.4.7 Drug-target network

Cytoscape 2.6.0 [190] was used to generate the network graph. Known interactions shown as pink edges were collected from DrugBank annotations and literature search. For example gefitinib-MAPK10 and gefitinib-MAPK-14 are not annotated expressly by DrugBank, but gefitinib has been shown to inhibit MAPK14 with an IC_{50} of 1.19μM [43] and MAPK10 with an IC_{50} of 2.3μM [42] through two different kinase binding assay studies.

3.4.8 Cell lines and reagents

The SUM149 cell line was purchased from Astrand (Ann Arbor, Michigan, USA) and grown according to the manufacturer's recommendation. In brief, cells were cultured in F-12 (Ham's) media (Gibco/Invitrogen, Burlington, Ontario, USA) supplemented with 5 μg/ml insulin (Sigma, Oakville, Ontario, Canada) 1 μg/ml hydrocortisone (Sigma), 10 mM HEPES (Sigma), and 5% fetal bovine serum (Gibco/Invitrogen). BT474-m1 cells (50% F12/50% DMEM) were obtained from MC Hung, M. D. Anderson Cancer Center. MDA-MB-231 (DMEM with 10% FBS) and HCC1937 (RPMI with 5% FBS) cells were from the American Type Culture Collection (ATCC). MDA-MB-435/LCC6 cells were a gift Dr. Robert Clark at Georgetown University, Washington, DC, and were cultured in DMEM (Invitrogen, Burlington, ON, Canada) with 2 mmol/L L-glutamine and 10% fetal bovine serum (Invitrogen)

3.4.9 Growth assays

Breast cancer cells were seeded in 96-well plates (5000 cells/well) and incubated for 24 hours at 37˚C. Cells were then treated with gefitinib (isolated from tablets purchased from AstraZeneca and kindly provided by Ching-Shih Chen (Ohio State University, USA) at 1μM and with TDF (≥97% pure tenofovir disoproxil fumarate powder purchased from Changzhou Huaren Chemical Co, Jiangsu, China) at 1μM, 10μM, and 100μM. Cells were also treated with vehicle controls: DMSO for gefitinib and methanol for TDF. Nuclei/cell counts were determined after 72 hours of drug treatment. Cells were washed in PBS and then fixed and stained in 2% paraformaldehyde containing Hoechst dye (1μg/ml). Cell numbers were determined using the ArrayScan VTI high throughput analyzer. Each cell count reported is the average of six wells. IC_{50} values were estimated by linear regression of a log-transformation of the data. Significant decreases in cell count were assessed using an unpaired Student's t-test p-value < 0.05.

Growth assays with gefitinib-insensitive cell lines MDA-MB-231 and HCC1937 were performed using Truvada (Gilead Sciences, Inc., a combination drug with TDF and emtricitabine). Truvada pills were crushed and completely dissolved in DMSO to create

stock solution. These assays were performed as the average of two wells (only one well in the case of HCC1937).

3.4.10 Western blotting

SUM149 cells were plated in 6-well plates with a density of $3.5x10^5$ cells/well and incubated for 24 hours at 37˚C. Cells were then treated with vehicles, gefitinib (1μM), and TDF (1 and 10μM) for 24 hours then lysed with ELB buffer (5mmol/l PH 7.4 HEPES, 150 mmol/l NaCl, 1 mmol/L pH 8 EDTA, 1% Triton X-100, 1% sodium deoxycholate, and 0.1% SDS) with protease inhibitors. Proteins were then quantified by Bradford assay, resolved on a 12% SDS-PAGE, and transferred overnight to nitrocellulose membranes at 40V and 4˚C. Membranes were blocked in 5% milk in TBS/0.1% Tween for 1 hour at room temperature, before being probed with primary antibodies as follows: p42/44 ERK (1:1000; cell signaling) phospho-ERK (1:500; cell signaling), EGFR (1:1000; stressgen), phospho-EGFR (1:1000, cell signaling), vinculin (1:1000; cell signaling), and incubated with either mouse (1:5000; Amersham) or rabbit (1:2000; Amersham) secondary antibodies. Protein bands were visualized using ECL Western blotting detection reagents (GE Healthcare).

In the EGF dependence experiment, SUM149 cells were serum-starved for 24 hours before treating with vehicles, gefitinib (1μM) and TDF at 1 and 10μM for a further 24 hours. Then one set of cells remained serum starved, whereas the second set was stimulated with EGF (20ng/mL) for 30 minutes. Western blots were then conducted as described above.

3.4.11 Direct binding assays

Caliper Discovery Alliances and Services performed an *in vitro* assay of TDF against EGFR using their LabChip 3000 technology. An enzyme titration was performed in standardized buffer with 100μM ATP, 1μM fluorescent peptide substrate, and the drug on a microfluidic chip. A capillary sipper was used to separate the phosphorylated and unphosphorylated substrate via electrophoresis. The amount of phosphorylated substrate is then measured via laser-induced fluorescence. It is reputed to be able to detect both strong and weak inhibitors due to the sensitivity of microfluidic precision and can identify drug candidates missed by conventional techniques. TDF was tested at 8 concentrations in duplicate.

Invitrogen's SelectScreen service performed an *in vitro* assay of TDF against EGFR using their Z'-LYTE methodology. This assay uses a synthetic peptide substrate with a fluorophore on each end (the two fluorophores have different emission wavelengths). A site-specific protease was then used to cleave only non-phosphorylated peptides, disrupting the fluorophore pair and affecting the ratio of the two fluorophore emissions. This difference of emission was measured. TDF was tested at 10 concentrations in duplicate.

SignalChem (Richmond, BC, Canada) performed in an *in vitro* assay of TDF against EGFR. Kinase assays consisted of ^{33}P-ATP at 5μM, the protein kinase, peptide substrate, assay buffer, and the drug. After 20 minutes of incubation, the radiolabelled substrate was isolated using phosphocellulose paper and the fluorescence reading was obtained. Blank assays without substrate or drug, and assays without the drug, were used as controls. Staurosporine at 1μM was used as the positive control drug.

3.4.12 Alignment of the ERBB kinase family

Protein sequences of the four ERBB family kinases were downloaded from UniProt [191] and aligned using standard settings in the ICM graphical user interface [87]. The ICM method is a global, Needleman-Wunsch alignment with zero end gap penalties [355]. The residue substitution matrix and normalized gap penalties were determined by training on a fold-recognition benchmark [355]. In method comparison studies, the ICM method has performed well, particularly in its ability to allow continuous gaps in both sequences being aligned [356]. This is significant for proteins because two proteins sharing a fold will have conserved regions but will also have regions (such as loops) with significant structural deviation. The multiple sequence alignment (MSA) algorithm is based on CLUSTAL, constructing a guide tree to cluster sequences and building the MSA by aligning proteins from most to least similar [357]. The kinase domain sequence was determined using the sequence of EGFR kinase domain crystal structure 1M14.

Figure 3.1 Predicted binding sites in EGFR crystal structures.

a) The three largest predicted binding sites the EGFR kinase domains. b) The three largest predicted binding sites in the EGFR extracellular domain. The protein is shown in green with the binding pockets shown as green, orange, and blue in order of decreasing size. In total, the 23 EGFR crystal structures had 65 predicted binding pockets.

(a) (b)

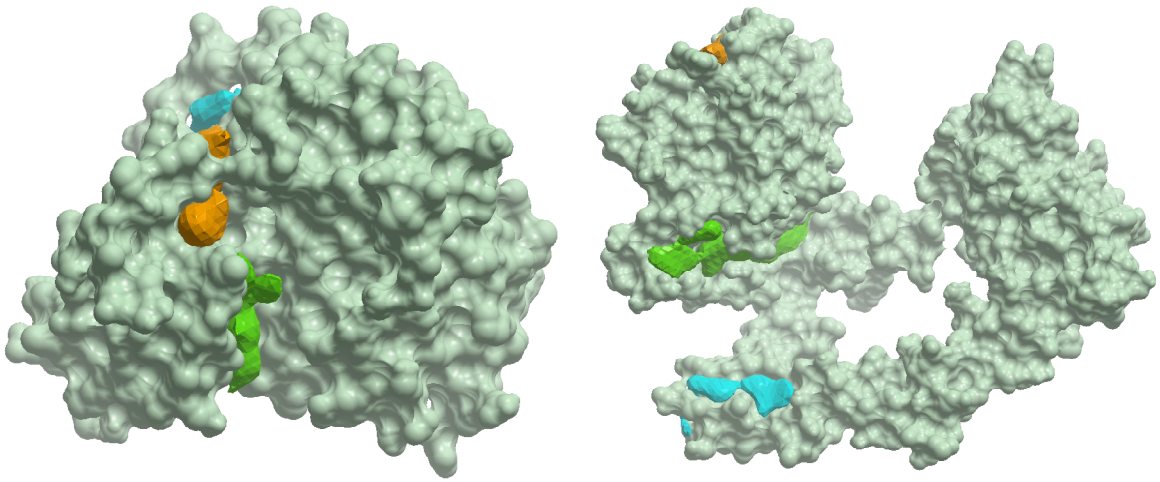
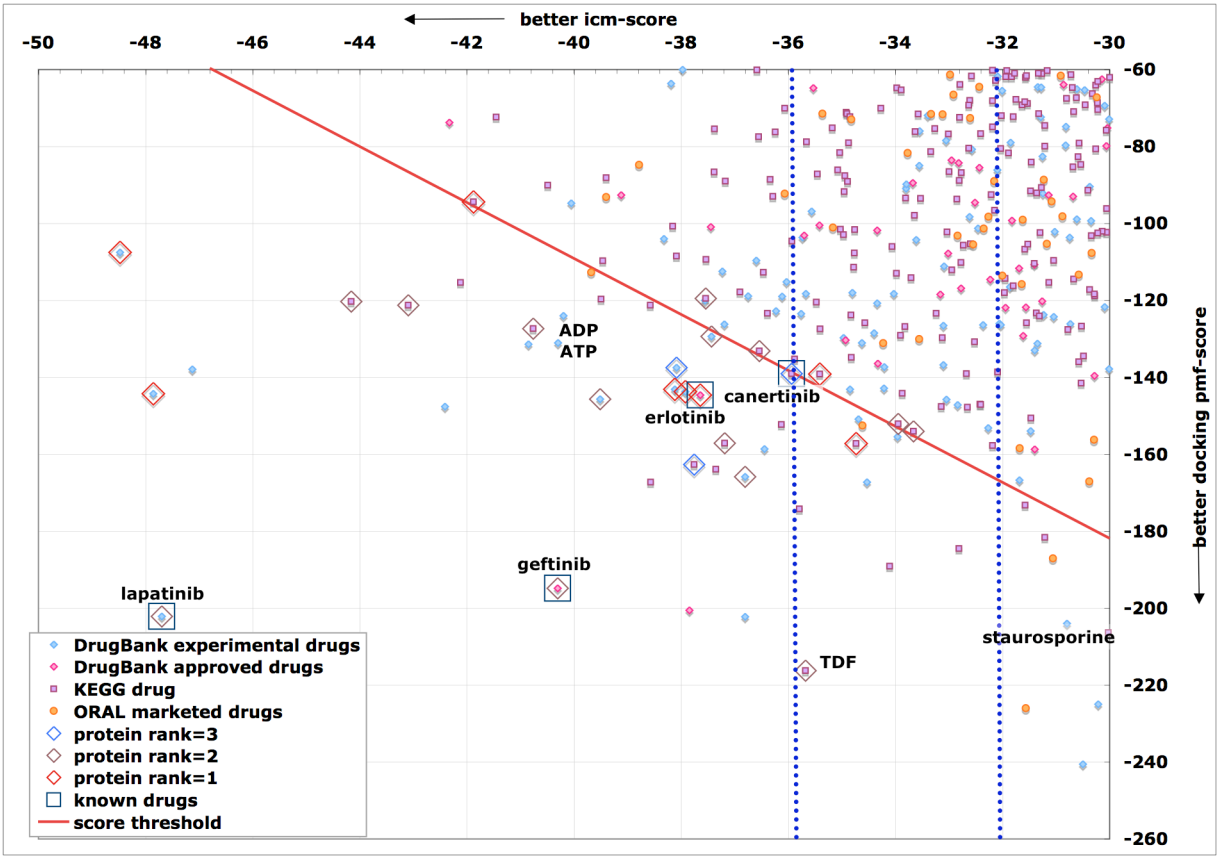


Figure 3.2 Molecular docking analysis.

(a) Chemicals from DrugBank, KEGG drug, and an oral-marketed drug database [208] were docked to EGFR crystal structures. Docking scores (icm-score) and potential of mean force scores (pmf-score) were plotted for each EGFR-drug interaction. (b) Predicted binding conformation of two known EGFR ligands (erlotinib and ATP) compared to their native EGFR-ligand crystal structures. The solved inhibitor is shown in grey stick format, and the neighbouring protein residues are in white sticks. The docked inhibitor is shown as yellow sticks. All compounds exhibited RMSD values under 2.0Å as expected for a successful docking result.



b) Erlotinib. RMSD 1.1Å. ATP. RMSD 1.2Å.

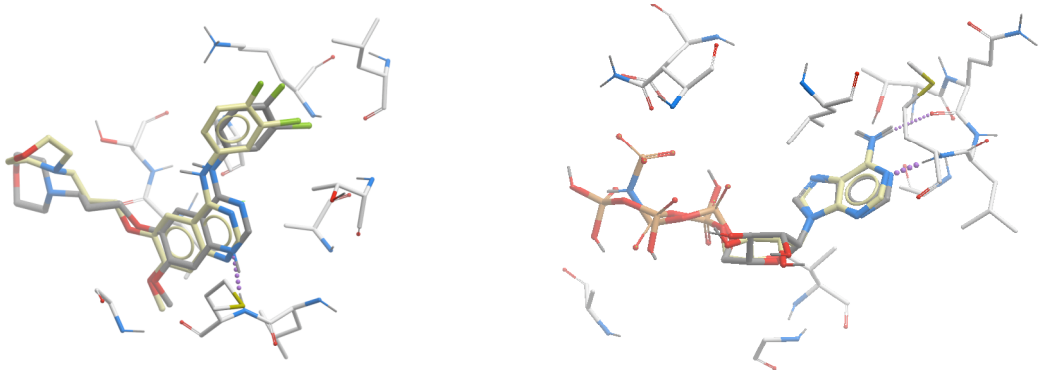


Figure 3.3 Drug-target interaction network of predicted EGFR drugs.

Targets that are connected to EGFR through one drug interaction are shown. Targets are depicted as boxes, and drugs as circles. Approved drugs are purple and experimental drugs are shown by their DrugBank ID in blue. Predicted interactions are depicted as edges, with literature-verified interactions colored pink. The width of the edge line shows the interaction's protein rank (thickest line = protein rank 1, thinnest line = protein rank 3). Only interactions with protein rank ≤3 are shown.

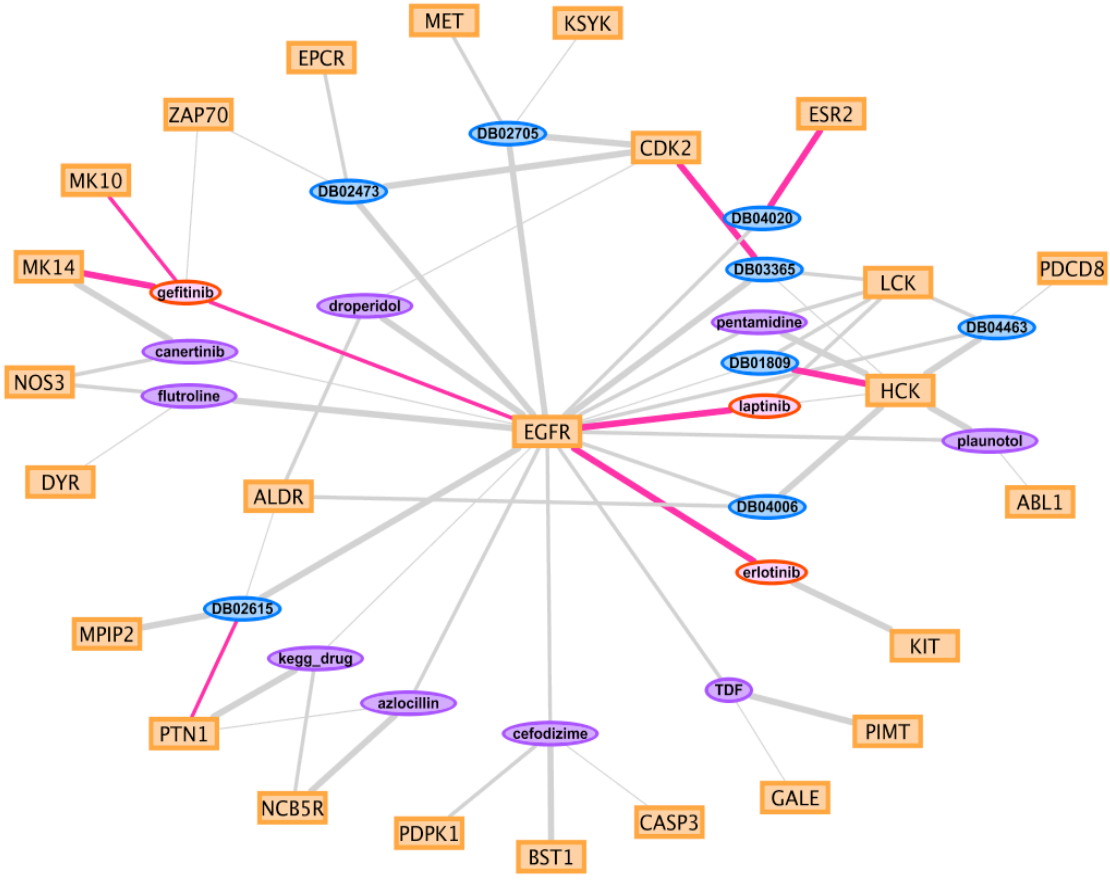
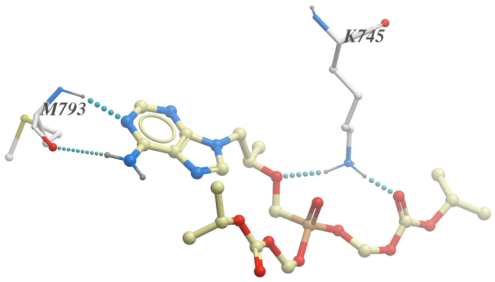


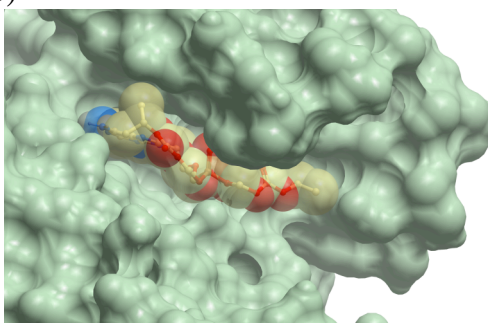
Figure 3.4 TDF is predicted to bind to the ATP-binding site of EGFR

The predicted binding mode of TDF in the ATP binding pocket of EGFR is shown as (a) a stick model and (b) a space-filing model. (c) Comparison of binding modes of TDF and an ATP-analog (grey stick). Hydrogen bonds formed by TDF are in purple and those by the ATP-analog are in green.

(a)



(b)



(c)

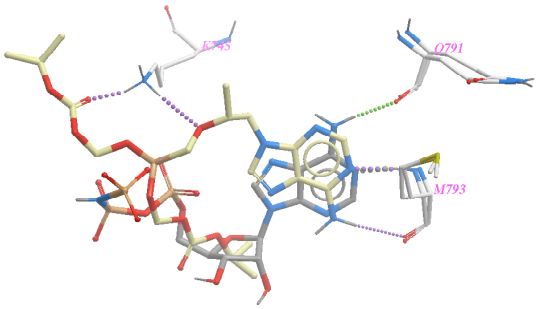
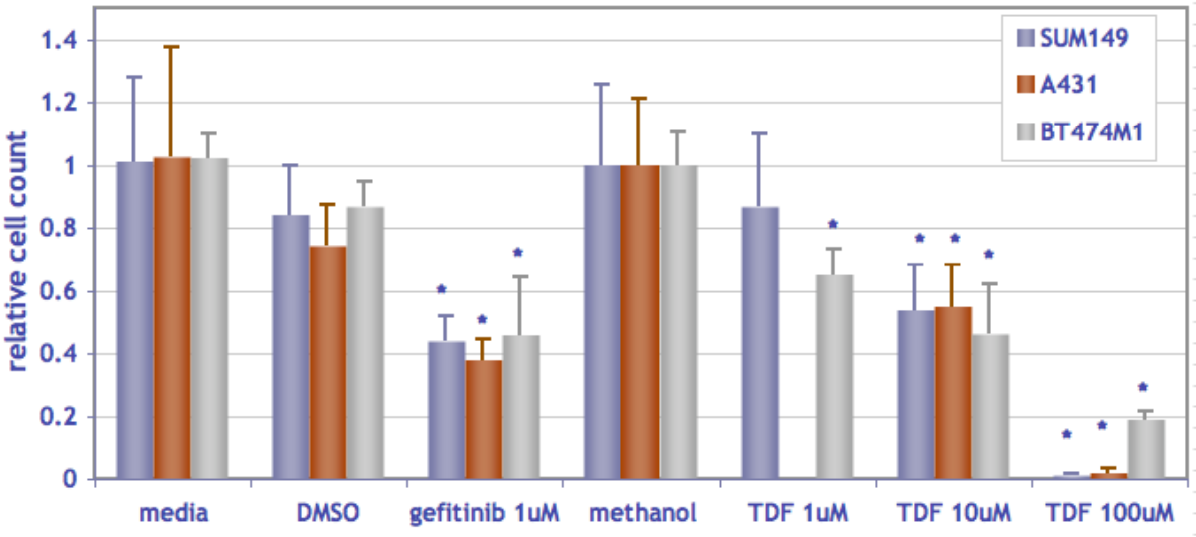
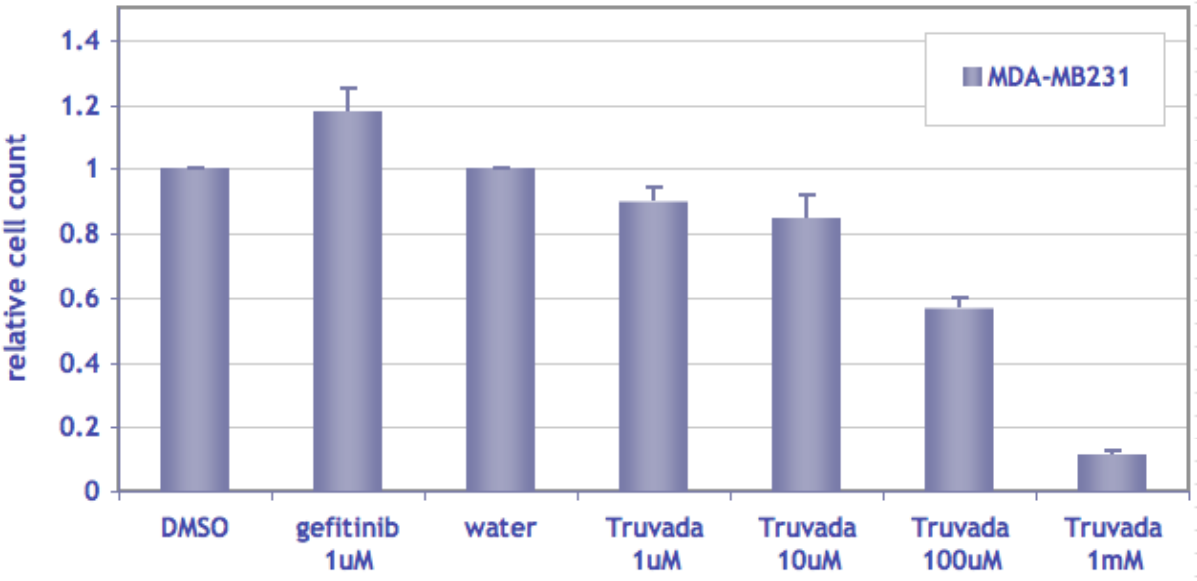


Figure 3.5 TDF suppresses the growth of gefitinib-sensitive cancer cell lines only.

a) SUM149 and A431 cells are known to overexpress EGFR. DMSO was the vehicle control for Iressa (gefitinib) and methanol was the control for the drug TDF. BT474M1 expresses EGFR but is also known to overexpress HER2. Decreases in growth that are statistically significant are indicated with * (p-value < 0.05).



b) The MDA-MB231 breast cancer cell line is not sensitive to gefitinib. Both gefitinib and TDF did not inhibit this cell line in my study. The drug used here was Truvada (TDF + emtricitabine) obtained from dissolving crushed pills. DMSO was the vehicle control for gefitinib and methanol was the control for Truvada.



c) LCC6 is known to overexpress HER2, and has been shown to be sensitive to Iressa under low serum conditions and insensitive with 10% FBS [220]. In my standard assay protocol with 5% FBS, LCC6 did not show sensitivity to Iressa. However, LCC6 was inhibited in a dose-dependant manner by the drug TDF. DMSO was the vehicle control for gefitinib and methanol was the control for TDF.

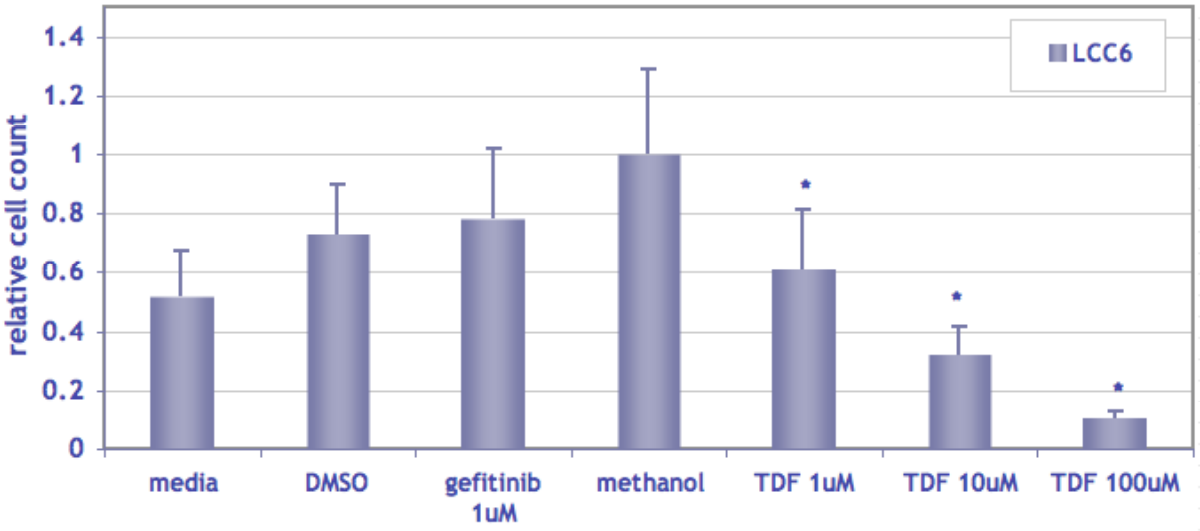
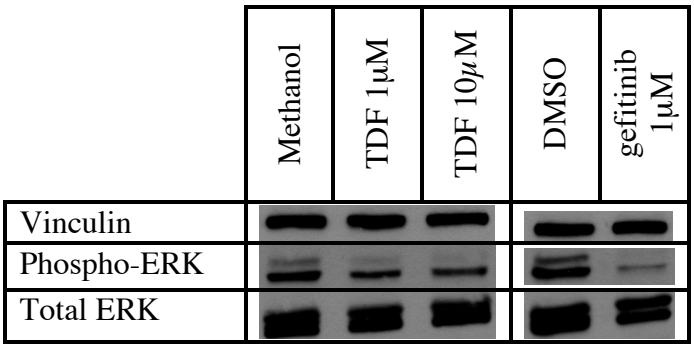


Figure 3.6 TDF inhibits EGFR pathway signaling.

(a) Effect of TDF on EGFR and the ERK signaling pathway. SUM149 cells were treated with two concentrations of TDF for 24 hours. Cells treated with TDF show a decrease in phospho-ERK signaling without decreasing total-ERK. Vinculin was used as loading control.



(b) Effect of TDF on ERK pathway signaling with and without EGF. SUM149 cells were serum-starved, treated with two concentrations of TDF for 24 hours, and then stimulated with EGF for 30 min (right). Gefitinib was used as a control drug (left). Vinculin was used as the loading control.

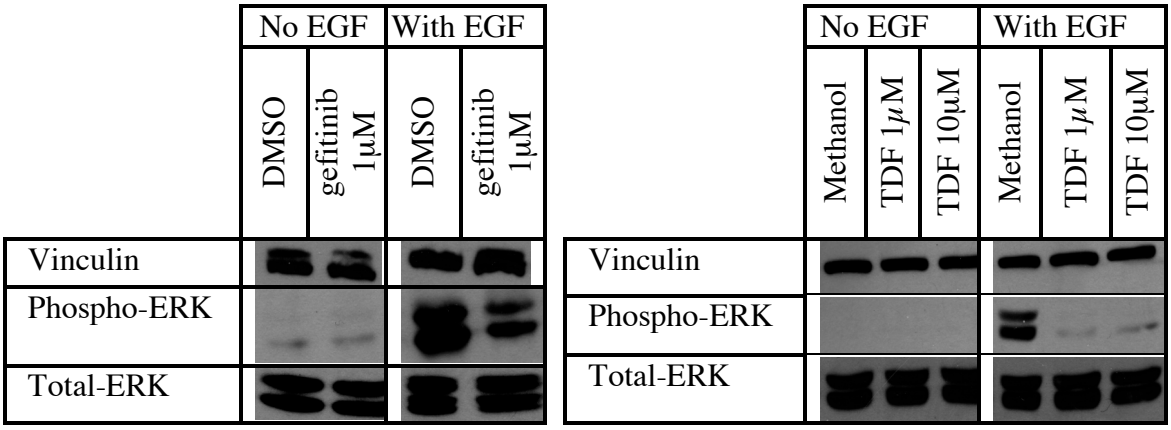


Figure 3.7 *In vitro* kinase assays of TDF against EGFR
a) *In vitro* kinase assay of control drug gefitinib. The known inhibitor gefitinib did not exhibit any EGFR inhibition (negative values represent % inhibition) at 50μM ATP concentration, however, adjusting the concentration to 5μM showed a strong dose-dependent response.

[ATP]	drug 10nM	drug 100nM	drug 1μM	drug 10μM	drug 100μM
5μM	-6	-17	-23	-73	-51
50μM	-99	-102	-103	-103	-103

b) *In vitro* kinase assay of TDF against EGFR at 5μM ATP concentration. The assay was performed by SignalChem (Richmond, BC, Canada) three times with various TDF concentrations. Overall, TDF did not appear to inhibit EGFR. Negative values represent % inhibition, and anything within +20 or -20 may represent noise.

[ATP]	drug 10nM	drug 100nM	drug 1μM	drug 10μM	drug 100μM
5μM	-40	-38	-4	-3	-
5μM	-3	3	-	-	-
5μM	-6	-	0	-6	-23

c) *In vitro* kinase assay of TDF against EGFR performed by Invitrogen and Caliper, respectively.

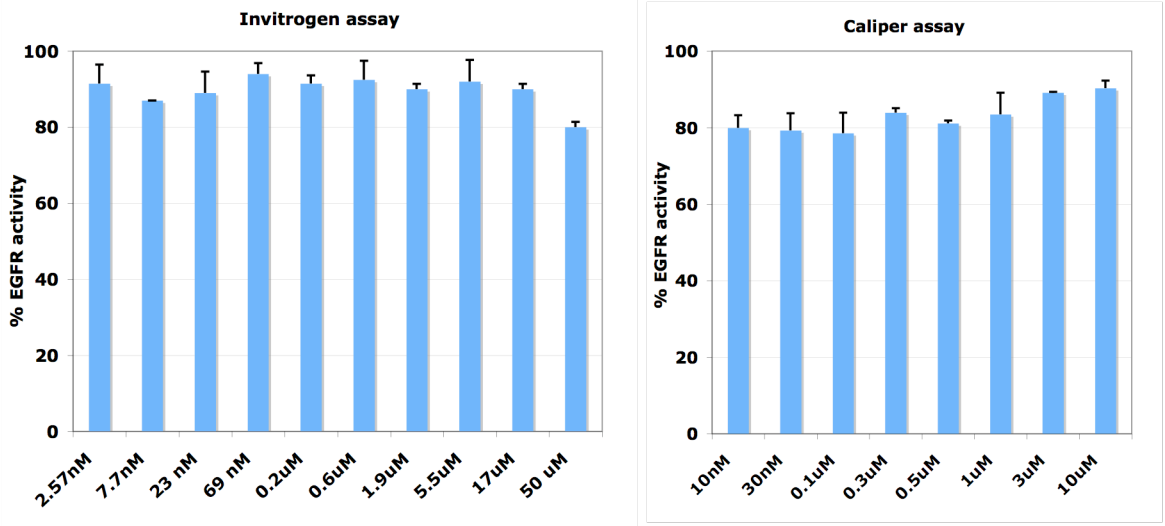


Figure 3.8 Comparison of the ERBB family of kinases.

Within the kinase domain, the four proteins are quite similar as shown in the below sequence alignment. They are colored by conservation (dark green is perfectly conserved, light green is conserved across 3 of the 4 kinases). Residues within 3.5Å of the ATP binding site are shown in red boxes and are quite well conserved, particularly across EGFR, ERBB2/HER2, and ERBB4/HER4.

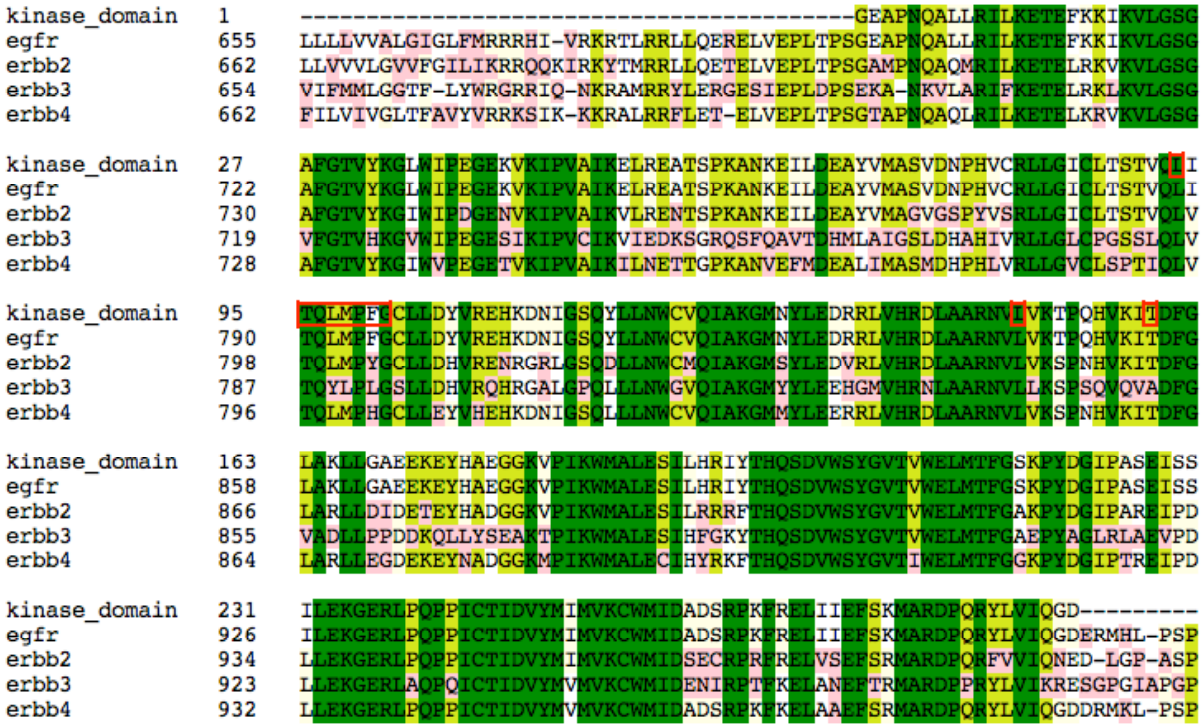


Table 3.1 The 24 crystal structures of EGFR and their mutational status.

PDB code	Domain	Ligand	Resolution (Å)	Mutation	Reference
1ivo	extracellular	-	3.30	-	[227]
1mox	extracellular	TGF-alpha	2.50	-	[228]
1nql	extracellular	EGF	2.80	-	[229]
1yy9	extracellular	cetuximab	2.61	-	[230]
1m14	kinase	-	2.60	-	[231]
1m17	kinase	erlotinib	2.60	-	
1xkk	kinase	lapatinib	2.40	-	[232]
2gs2	kinase	-	2.80	-	[233]
2gs6	kinase	ATP analog	2.60	-	
2gs7	kinase	AMP-PNP	2.60	-	
2itn	kinase	AMP-PNP	2.47	G719S	[234]
2ito	kinase	gefitinib	3.25	G719S	
2itp	kinase	AEE788	2.74	G719S	
2itq	kinase	AFN941	2.68	G719S	
2itt	kinase	AEE788	2.73	L858R	
2itu	kinase	AFN941	2.80	L858R	
2itv	kinase	AMP-PNP	2.47	L858R	
2itz	kinase	gefitinib	2.80	L858R	
2itw	kinase	AFN941	2.88	-	
2itx	kinase	AMP-PNP	2.98	-	
2ity	kinase	gefitinib	3.42	-	
2j6m	kinase	AEE788	3.10	-	
2j5e	kinase	13-jab	3.10	-	[235]
2j5f	kinase	34-jab	3.00	-	

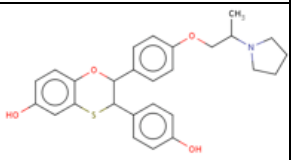
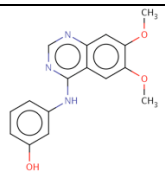
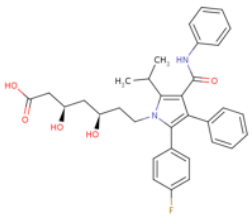
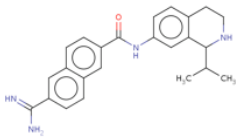
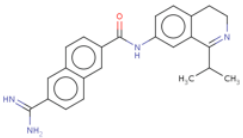
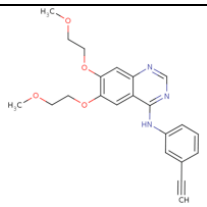
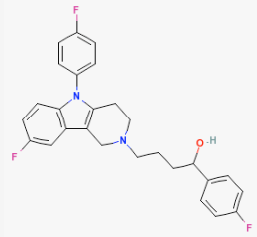
Table 3.2 Various scoring thresholds for the docking results.

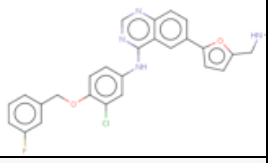
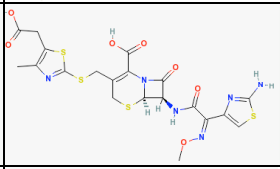
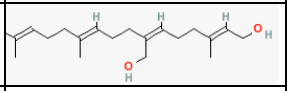
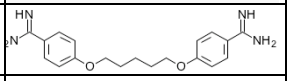
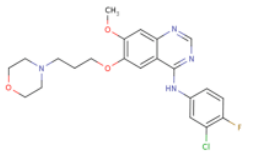
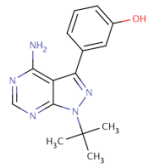
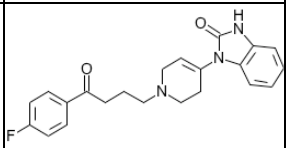
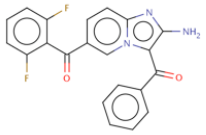
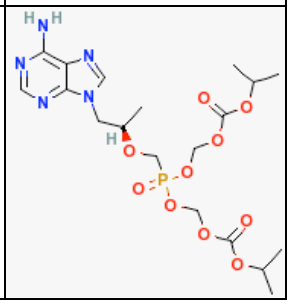
Compared to the default score threshold, the devised consensus line predicts far fewer EGFR inhibitors out of the 5908 drugs, while retaining the ability to predict known EGFR ligands. The protein rank filter only predicts 4 known ligands, but the eliminated 3 predicted inhibitors are broad-spectrum kinase ligands (ATP, ADP, and staurosporine).

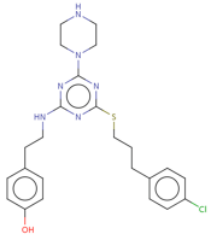
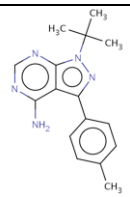
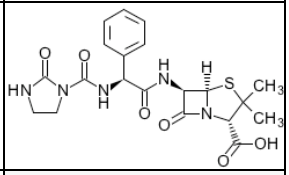
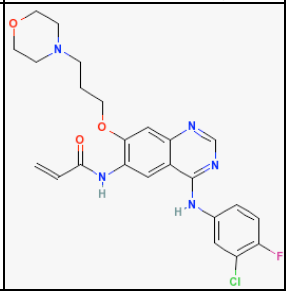
Score cut-off method	Number of predicted inhibitors	Number of known ligands	Enrichment factor versus random
No threshold (random selection)	5903	7	1
Default score threshold -32	528	7	11
icm-score -32, pmf- score -60 minimum threshold	226	7	26
Use known-drug score -36	150	7	39
Use consensus scoring threshold line	50	7	118
Use EGFR protein rank of 3 or lower from docking the top compounds against a panel of 134 drugs	20	4	169

Table 3.3 Top 20 known drugs predicted to inhibit EGFR.

The table is sorted by the protein rank (prot rank) of EGFR for each drug, then by icm-score. These interactions passed the consensus score filter and have a protein rank of at least 3. The majority of drugs are predicted to bind the wild-type ATP-binding site; compounds 7 and 9 are exceptions, as they are predicted to bind to a site in the extracellular domain. Bolded compounds indicate known EGFR inhibitors

No.	icm score	pmf score	prot rank	No. of targets	Drug name	Experimental drug structures
1	-46	-107	1	13	compound 19	
2	-48	-144	1	9	4-[3-hydroxyanilino]-6,7-dimethoxyquinazoline	
3	-42	-94	1	10	Atorvastatin	
4	-38	-143	1	6	6-[n-(1-isopropyl-1,2,3,4-tetrahydro-7-isoquinolinyl)carbamyl]-2-naphthalene carboxamidine	
5	-38	-144	1	7	6-[n-(1-isopropyl-3,4-dihydro-7-isoquinolinyl) carbamyl]-2-naphthalene carboxamidine	
6	**-38**	**-145**	**1**	**2**	**Erlotinib**	
7	-35	-157	1	3	Flutroline	

No.	icm score	pmf score	prot rank	No. of targets	Drug name	Experimental drug structures
8	**-48**	**-202**	**2**	**6**	**Lapatinib**	
9	-44	-120	2	16	Cefodizime sodium	
10	-43	-121	2	9	Plaunotol	
11	-41	-127	2	6	Pentamidine isetionate	
12	**-40**	**-195**	**2**	**9**	**Gefitinib**	
13	-40	-146	2	4	3-(4-amino-1-tert-butyl-1h-pyrazolo[3,4-d] pyrimidin-3-yl) phenol	
14	-37	-157	2	6	Droperidol	
15	-37	-129	2	9	[2-amino-6-(2,6-difluoro-benzoyl)-imidazo[1,2-a]pyridin-3-yl]-phenyl-methanone	
16	-36	-216	2	3	Tenofovir disoproxil fumarate	

No.	icm score	pmf score	prot rank	No. of targets	Drug name	Experimental drug structures
17	-36	-166	2	2	4-(2-{[4-{[3-(4-chloro phenyl) propyl]sulfanyl}-6-(1-piperazinyl)-1,3,5-triazin-2-yl]amino}ethyl) phenol	
18	-38	-137	3	4	1-ter-butyl-3-p-tolyl-1h-pyrazolo[3,4-d]pyrimidin-4-ylamine	
19	-38	-163	3	5	Azlocillin sodium	
20	**-36**	**-139**	**3**	**6**	**Canertinib**	

4 Combining Virtual and High-throughput Screening to Discover Novel Repositioning Candidates for Triple Negative Breast Cancer

4.1 Introduction

Breast cancer tumors are notorious for their phenotypic and genetic diversity. Seminal papers in the early 2000's used gene expression profiling to classify breast tumor samples by gene expression profiling, and found four to six major subtypes [236, 237]. Clinically, three major receptors are used for molecular classification of these tumors: the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER-2). ER- or PR- positive tumors represent over 50% of breast cancer tumors [238] and have long been treated using hormonal therapies like tamoxifen [239]. 15-25% of early stage breast cancers overexpress HER2 and are treated with the anti-HER2 monoclonal antibody trastuzumab (Herceptin) [240]. 10-20% of breast cancers are triple-negative (TNBC) – characterized by a lack of ER, PR and HER2 expression – and have been associated with aggressive clinical course and poor prognosis [241]. Despite being only a relatively small proportion of breast cancers, TNBCs have a significantly higher rate of death within three years following diagnosis: in a large cohort of 1,601 women, the median time to death was 4.2 years for TNBC patients compared to 6 years for other types of breast cancer [241]. Also unlike the other two subtypes, TNBCs do not have an established therapeutic target or targeted therapies for clinical use. Instead, they are treated with standard chemotherapy regimens and have the highest rates of relapse and metastasis [242].

As TNBCs do not express the ER, PR and HER2 receptors, it has been suggested that they may be more responsive to non-receptor mediated therapies [243]. One protein that has emerged as a potential drug target in TNBC is the p90 ribosomal S6 kinase (RSK), a member of the MAPK pathway. Following growth factor stimulation, MAPK signaling proceeds through RAS, RAF and ERK. ERK activates RSK, which in turn phosphorylates a host of downstream substrates involved in nuclear signaling, cell cycle progression, cell survival, motility, and proliferation [244]. RSK is also the predominant activator of the Y-box binding protein-1 (YB-1) transcription factor [245]. Upon phosphorylation at the S102 site, P-YB-1-

S102 migrates to the nucleus and turns on transcription of many genes important in cancer proliferation, such as EGFR [246]. In a study of 48 cancerous and 12 normal breast tissue samples tissue samples, 50% of the cancer tissues had RSK1 or RSK2 isoform overexpression [247].

In this study, I used molecular docking to virtually screen (VS) the Prestwick Chemical Library of 1,120 off-patent drugs, in order to identify novel repositioning candidates for the RSK1 and RSK2 proteins. By analyzing known RSK inhibitors, I was able to formulate stringent scoring, ranking, and visual criteria to select the Prestwick drugs most likely to inhibit RSK activity. We also performed high-throughput screening (HTS) of the Prestwick library against RSK, resulting in a second and significantly different, list of RSK inhibitors. Top candidates that showed strong HTS RSK inhibition and strong VS dockings were validated in secondary experimental screens. The ability of these drugs to inhibit YB-1 phosphorylation and nuclear translocation was confirmed, as well as their ability to inhibit growth of TNBC cell lines. Overall, three Prestwick drugs were validated as novel RSK inhibitors and potential TNBC repositioning treatment strategies.

4.2 Results

4.2.1 Building RSK models to supplement existing RSK structures

RSK proteins contain two kinase domains connected by a linker region. The C-terminal domain (CTKD) autophosphorylates the N-terminal domain (NTKD), which is required for RSK activation. Subsequently, the NTKD phosphorylates downstream substrates such as YB-1 and GSK3B [248].

I gathered existing RSK crystal structures from PDB including NTKD and CTKD domains from RSK1 and RSK2 (Table 4.1). While these structures had acceptable resolutions for docking purposes (under 2.5Å), they also had shortcomings. First, there were several large structure gaps proximal to the known ligand binding sites in the NTKD structures (Figure 4.1a – shaded residues); docking small molecules next to gaps may result in false positive predictions. Second, none of the existing structures were solved in complex with a peptide substrate and thus were not in a substrate-binding conformation for rigid-protein docking. I

therefore decided to supplement the existing protein structures with two RSK1 NTKD homology models.

4.2.1.1 Creating a model of RSK based on ligand-binding and loop modeling

The first model of RSK1 NTKD utilized the RSK2 NTKD structure as a template (Figure 4.2). This resulted in a high confidence model based on 90% sequence identity between the two proteins' kinase domains. The RSK2 structure was complexed to an ATP analog called AMP-PNP without metal ion catalysts present in the binding site. It differed from the existing RSK1-AMP-PCP structure since metal ions often pull protein residues towards them, and the lack of metal ions in the RSK2 structure resulted in a dissimilar protein conformation.

There were two gaps in the template structure (Figure 4.1b) corresponding to highly flexible loops in the RSK protein. To construct loops during the homology modeling process, ICM refers to a loop database assembled from existing PDB structures, and then performs energy minimization based on its Monte Carlo method [249]. The resulting model (Figure 4.2a) shows that the presence of the activation loop (shown in grey) created a putative binding pocket that overlapped with the substrate binding site (substrate peptide position shown in orange).

4.2.1.2 Creating a model of RSK bound to YB-1 peptide

I also built a homology model of RSK1 NTKD based on a peptide-bound AKT structure. Since it is known that RSK and AKT can both phosphorylate GSK3B and YB1 at a serine residue (Figure 4.3a), I hypothesized that the binding conformations of these interactions would be similar. In the RSK-AKT sequence alignment (Figure 4.1b), it can be seen that the AKT template structure 1o6k does not have gaps involving loop regions; I was thus able to confidently homology model the loop positions in a peptide-bound protein conformation.

The final model is shown in Figure 4.3 (b,c). The protein-peptide interaction is stabilized by 11 hydrogen bonds along the peptide length. Comparison of the predicted RSK-YB-1 binding conformation with the solved structure of AKT-GSK3B (Fig 4.3 c,d) shows

similarities in hydrogen bonds and interacting protein residues. The side chain R2 of the YB-1 peptide forms three hydrogen bonds with RSK (E254, Y228, E191) and these bonds are retained in the GSK3B-AKT complex. The peptide (catalytic) serine is in a position to hydrogen bond to the protein as well as receive a phosphate group from ATP. In addition, the interactions between the peptide and both G244 and K189 are consistent. The peptide binding pocket predicted in this second model was entirely different from the pocket in 4.2.1.1, due to the change in loop position (Figure 4.2b).

4.2.2 VS off-patent drugs against RSK

I used my established molecular docking pipeline to search for existing drugs that could potentially inhibit RSK. First, I collected existing PDB structures of RSK as well as the two modeled structures. Both the NTKD and CTKD of RSK1 and RSK2 were represented in my binding site database. The NTKDs are very similar between the two RSK isoforms, with 90% sequence identity; in contrast, the NTKD and CTKD of each RSK protein are very different, with only 27% identity. Thus, for this analysis I categorized the binding sites as N-terminal or C-terminal, without regard for the RSK isoform.

For each RSK protein, the ATP- and substrate- binding sites were prepared for docking, resulting in 16 binding pockets. As described in previous chapters, each drug is docked to all pockets, and the best site-drug score becomes the representative RSK-drug score. The results from docking 1,120 Prestwick compounds and 10 known RSK inhibitors are shown in a score plot (Figure 4.4) and are discussed in more detail in the following sections. Analogous to previous chapters, I used a consensus score threshold and protein ranks to aid in removing potential false positive interactions.

4.2.2.1 Docking known inhibitors to RSK

To assess my VS strategy, I first docked known small molecules inhibitors of RSK to my pocket database. There were ten in total, including four compounds from existing RSK structure complexes and six chemically diverse RSK inhibitors (in vitro IC_{50}'s 15nM to 1μM). The docked conformations and scores for these known binders are shown in Table 4.2.

4.2.2.1.1 Known inhibitors had good docking scores and conformations

Four compounds were present in RSK structure complexes and thus represented 'cognate' docking scenarios (Table 4.2 a-d). Cognate scenarios are less challenging to dock well, since the protein is already in the ligand-bound conformation. Aside from the ATP analog AMP-PCP, the other eight inhibitors showed icm-scores ranging from -29 to -54 and pmf-scores ranging from -102 to -190. These values were indicative of potential binding interactions as they were close to or better than the icm-score cut off of -30 established in Chapter 2. There was no correlation between the strength of the icm-score and the strength of RSK inhibition, in agreement with previous studies [109]. However I did note a weak correlation between pmf-score and binding strength (r-squared value from linear regression 0.67).

The predicted binding conformations of AMP-PNP, AMP-PCP, purvalanol A, and staurosporine were compared to their respective PDB structure complexes (Table 4.2 a-d, grey versus yellow compounds). They exhibited RMSD values from 0.3-2.2Å, which indicated an accurate prediction of the bound conformation. In addition, hydrogen bond contacts in the original PDB complexes were retained in the predicted conformations. The remaining six compounds were not solved in a PDB RSK complex and thus represented the more challenging computational scenario of non-cognate docking. However, it can be seen in Table 4.2 (e-j) that they all exhibited icm-scores indicative of likely binding interactions, ranging from -28 to -36. Further examination of protein hydrogen bonding residues suggested that important residues for small molecule inhibitor binding were Q70, S72, F73, Q74, D142, L144 for the NTKD and C436, E463, T493, E494, M496, D561 for the CTKD.

4.2.2.1.2 Known inhibitors docked well to multiple RSK structures

This score table in Table 4.3 shows the icm-score of known inhibitors docked to each of the RSK binding sites. The best scoring structure for each compound is shaded blue and cognate docking scenarios are boxed in black. The distribution of good scoring interactions (shaded orange) across the table shows that each structure was able to dock a different set of known inhibitors well. The exception was 2z7q, which did not dock any known inhibitors with acceptable icm-scores. 2z7q was also the only cognate docking scenario with a poor score,

suggesting that this PDB entry encountered errors during ICM protein preparation steps. Thus, 2z7q was removed from further analysis. Overall, I found that including multiple structures of the ATP-binding domains was integral to the docking analysis as each of the structures had characteristics for binding different known inhibitors.

4.2.2.1.3 Known inhibitors docked well to specific kinase domains

For four of the known inhibitors, the exact binding domain of binding has been elucidated. sl0101 binds to the RSK NTKD, since a mutant RSK with a modified NTKD (replacing the 12 amino acid loop involved in ATP binding of RSK (p90RSK) with that of p70RSK) was much less inhibited [247]. Correspondingly, I found that sl0101 docked well to several NTKD structures but to none of the CTKD structures (Table 4.3j). fmk inhibits RSK2 though its electrophilic fluoromethylketone moiety interacting with Cys436 of RSK2 CTKD [250]; this is reflected in the docked binding conformation (Table 4.2e) as well as the fact that it docked well it all three CTKDs but only one of six NTKDs. nsc356821 was discovered through a VS against the RSK2 NTKD [251]; here, its best scoring structure when docking is a NTKD but it also intriguingly docks well to the CTKD structures (Table 4.3j). BI-D1870 binds at the NTKD of all RSK isoforms, and can still inhibit a CTKD lacking mutant of RSK [252]. My results agree - as seen in Table 4.3e, BI-D1870 docks well to three NTKDs but to none of the CTKDs. In short, docking was able to discriminate between N-terminal and C-terminal binders.

4.2.2.1.4 Many top predicted interactions for known inhibitors were validated

Lastly, I examined whether the scoring and ranking criteria could predict novel true binding interactions. I determined the protein ranks of the ten known inhibitors by docking them to the 252 drug-target structure database constructed in Chapter 2. This metric measured whether a drug docked better to RSK compared to other drug targets. In Table 4.4, I list the top 21 scoring and ranking interactions (both in terms of icm-score and pmf-score) out of 2520 potential interactions. Literature search supported or validated 15 of these interactions (Table 4.4 references column). In addition, the top predictions for ATP analogs (AMP-PCP, AMP-PNP) and the pan-kinase inhibitor staurosporine were all kinases, despite the target

database consisting of a wide array of drug-target types. This extra validation step confirmed that the ranking criteria were able to aid in selecting true binding interactions.

Overall, the known inhibitor analyses established that my docking, scoring, and thresholding methods could dock and score known binders well, discriminate C-terminal inhibitors from N-terminal inhibitors, and predict true binding interactions. This result gave us confidence to perform further VS with different compound sets.

4.2.2.2 Top predicted inhibitors of the RSK ATP-binding site

The Prestwick collection of 1,120 off-patent drugs was docked to each of the RSK binding sites. The icm- and pmf- scores for the drugs are shown in Figure 4.4. As expected, known inhibitors (red boxes) landed in the bottom left-area of the plot with strong icm- and pmf-scores.

A ~10% consensus threshold (orange line) was applied to isolate the top scoring drugs in the bottom-left area of the plot. These top 110 drugs were docked to the 252 drug targets; for each drug, the rank of the RSK score compared to other proteins determined the protein rank. However, there were only 9 drugs with protein rank under 3 (which is the threshold I used in previous chapters); instead, I relaxed the thresholds to a protein rank of 10, selecting 27 drugs.

I did not anticipate a large number of RSK binders in the Prestwick collection of only 1,120 drugs, and thus did not want to eliminate potential true positive interactions using strict thresholds. Instead, I performed a visual inspection of drug binding conformations and eliminated top interactions not passing certain visual criteria. First, I selected drugs with hydrogen bonds spread out along the length of the bound drug conformation, which would better anchor the binding event. With this criterion, compounds such as picotamide were eliminated as it only interacted with RSK through on the left-side of the figure in Table 4.5e, while the isopthalamide moiety remained free. I note, however, that the known RSK inhibitors SB216763 and BI-D1870 only bound through one anchor point to L144. Thus,

eliminating drugs such as ellipticine (Table 4.5f) may be too stringent, since it also bound to RSK through a single hydrogen bond to L144.

Second, I checked that the residues interacting with the drug (i.e. the hydrogen-bond forming residues) overlapped with the residues previously determined from the known-drug dockings. As an example, isocixam (Table 4.5d) interacted with residues not involved in binding RSK-specific inhibitors, K94 and T204. Aside from L144, xamoterol (Table 4.5c) also bound to other residues: K94, R186, and T204. Etifenin (Table 4.5a) was eliminated as it also bound to F73 and K210 which did not seem to be involved in binding known inhibitors.

Third, I considered whether the compound was in a plausible docked conformation For example, an existing PDB structure of dobutamine bound to the beta-1 adrenergic receptor (2y00) exhibited a very linear conformation, unlike the predicted binding position (Table 4.5b). In contrast, methotrexate also had a 'kink' in its docking prediction, but binds in a conformation similar to an existing PDB structure (1rg7). Thus, methotrexate was not eliminated by this criterion.

Nine of the top 29 predicted RSK inhibitors – based on the consensus score, protein rank, and visual criteria - are summarized in Table 4.6.

4.2.2.3 Top predicted inhibitors of the RSK peptide-binding site

The score plot for screening the Prestwick library against the RSK peptide-binding site is shown in Figure 4.5. Overall, the compounds exhibited poor pmf-scores docked to the peptide pocket as compared to the ATP pocket. However, the distribution of scores still showed that the majority of compounds landed in a denser region of poor icm- and pmf-scores, while only a sparser set of compounds landed in the best-scoring bottom left of the score plot. Thus, consensus threshold criteria could still be applied to select the best icm- and pmf-scoring drugs from the set. After applying protein rank criteria, the top six predicted inhibitors of the peptide pocket are shown in Figure 4.5.

The lack of existing known RSK peptide-binding-site inhibitors to form comparisons upon also presented challenges during visual inspection of the top docking hits. However, I did know the residues important in binding the YB-1 and GSK3B substrates when building the protein model. I thus chose drugs that interacted with RSK residues such as S72, E191, E254, Y228, K189, D187 and G224 (Figure 4.3c).

The final list of five predicted RSK inhibitors are shown in Table 4.7, and comprise an entirely different set of compounds compared to the top ATP pocket predicted binders. All five appeared to bind in conformations that interfere with the peptide binding (Table 4.7: orange peptide in blue protein), and interacted with two or more RSK residues of interest.

4.2.3 HTS of off-patent drugs against RSK1

A high-throughput in vitro screen of Prestwick drugs against RSK1 using radiolabelled-ATP competition assays was conducted in parallel to the VS analysis. 32 of the 1,120 drugs exhibited significant inhibition of RSK at 10μM (>20% inhibition, Figure 4.6). It was expected that kaempferol would be a top HTS hit, since it is a substructure of the positive control drug sl0101 and a known inhibitor of RSK protein [247]. Kaempferol is a naturally occurring flavonol and other similar HTS hits included myricetin, hesperidin, luteolin, and apigenin. Several steroid hormones appeared in the list including estriol, progesterone, and estradiol-17-beta; however, previous studies have shown that the similar steroids estrogen and estradiol activated p90RSK through the MAPK signaling pathway [253, 254]. In the HTS assay, an agonist of RSK would have been detected by an increased fluorescence reading (higher levels of phosphorylated substrate). Thus, the contradictory inhibition of p90RSK activity observed suggested that the steroid hormone HTS hits may have been false positives.

4.2.4 Comparison of HTS and VS results

On the ATP pocket score plot (Figure 4.4), the ten known RSK inhibitors (red boxes) exhibited strong icm- and pmf- score pairs that landed in the bottom left of the plot. In contrast, the HTS hits (red points) did not separate as well from the remainder of the Prestwick drugs.

I determined the enrichment factors (EFs) of docking predictions for both the HTS-determined drugs and the previously known inhibitors (Table 4.8). Though the ten known ATP inhibitors were not part of the Prestwick collection, I assumed that they would have shown activity in the HTS screen and included them for a total of 1130 drugs. Compared to experimentally screening all compounds, VS threshold methods predicted interaction sets more enriched for true binders. For instance, applying a consensus score threshold and stringent protein rank criteria allows us to predict a set of 26 drugs containing six known ATP inhibitors. This resulted in a PPV of 23%, allowing us to select six of ten true RSK inhibitors without experimentally testing 98% of the compounds. However, the thresholds were not as effective at enriching for HTS hits, where only a 2.7x EF could be achieved.

The visual criteria performed better than stringent protein rank criteria with an EF of 7. Combining the visual criteria with loose protein rank criteria (using the worst protein ranks of the known inhibitors) resulted in the best EF of 11. I could not assess the EFs using visual criteria for the known RSK inhibitors, since their docked conformations were already the basis of the visual criteria.

Several of the HTS hits that did not pass score and rank thresholds are shown in Table 4.9. They generally had reasonable docked conformations and are predicted to interact with the previously determined 'important' RSK residue. The only drug in the set that would not have passed the visual criteria is menadione (Table 4.8 j), which formed just one hydrogen bond contact with the RSK protein.

To summarize, there were 32 HTS hits and 29 VS hits, with only 6 compounds shared in between the two lists. The known RSK inhibitors were more potent (IC_{50}'s 15nM to 1μM) and were better filtered using standard consensus score thresholds and protein-rank criteria, with an EF of 26; these criteria were not as effective for the HTS hits. By using score thresholds, visual criteria, and protein rank criteria, the VS method was able to predict a set 11x enriched for true binders with a 6/19=32% hit rate. This reveals that if I had used VS

without HTS, I would have been able to screen just 19 (1.7%) of 1,120 Prestwick drugs and discovered 6 novel RSK inhibitors.

4.2.5 Follow up validation of predicted hits

4.2.5.1 Secondary in vitro kinase screens

We validated the top HTS/VS predictions using low-throughput kinase assays against the RSK2 kinase. BI-D1870 was chosen as a positive control compound. Ellipticine and kaempferol were chosen as top scoring VS hits that also showed strong HTS inhibition. Menadione was selected as a lower scoring VS hit with strong HTS inhibition. Finally luteolin and apigenin were selected due to their plausible binding conformations despite weaker HTS inhibition. These five compounds were drug 'leads' with TNBC repositioning potential.

Figure 4.7 shows the activity of these 5 compounds when tested at low-throughput. Two substrate peptides were used to control for any effect the substrate might have on the inhibitor binding. Figure 4.7a shows the results using the YB-1 peptide that was also used in the HTS screen. The IC_{50} value of the positive control drug BI-D1870 was determined to be 16nM which was consistent with previous studies [252]. Overall, the IC_{50}'s were consistent with inhibition levels observed during the RSK1 HTS assay at 10μM drug concentrations. Menadione was a singular case as it only showed inhibition with one of the two peptides. Since it did not show strong inhibition of RSK-YB-1 in the secondary screen, it seemed likely that the HTS hit was a false positive. However, its ability to strongly inhibit the RSK-S6K interaction suggests that it still has potential as a RSK inhibitor.

4.2.5.2 YB-1 inhibition screens

RSK is the predominant kinase that phosphorylates YB-1 [245] and upon this event YB-1 is activated and translocates to the nucleus. Therefore, compounds inhibiting RSK would be expected to inhibit YB-1 activation and nuclear translocation. We therefore assessed the five lead compounds ability to affect nuclear phospho-YB-1 (P-YB-1) by western blotting (Figure 4.8a). As expected, the lead compounds all had decreased levels of P-YB-1 compared to YB-1, in contrast to the DMSO control. This result was shown both for the cytosolic and nuclear

fractions. In addition, immunofluorescence visualization with DAPI nuclear staining supported the decreased levels of P-YB1 in both the cytosol and nucleus (Figure 4.8b). Tthere were a wide range of potencies in the western blot results, in that ellipticine appeared to strongly decrease P-YB-1 but menadione did not. This result appears to agree with the secondary screen, where menadione poorly inhibited RSK in the presence of YB-1 substrate.

4.2.5.3 TNBC cellular growth screens

We further tested whether these RSK inhibitors were able to inhibit growth of two cellular models of TNBC: SUM149 and MDA-MB-231. For each cell line, both a monolayer and soft-agar colony growth assay was performed. The latter assay was especially important in order to better represent the 3D nature of the tumor and its environment.

BI-D1870 was once again used as the positive control drug and showed strong growth inhibition of both cell lines in both assays. We found that ellipticine and menadione significantly inhibited monolayer growth at 10μM, whereas kaempferol, phenindione and apigenin required higher concentrations to produce the same effect (Figure 4.9a). In anchorage-independent conditions, all five compounds inhibited TNBC growth at 10μM (Figure 4.9b). The drugs behaved similarly between the two cell lines, despite generally being less potent in the MDA-MB-231's. To ensure that these compounds were not cytotoxic, we tested them in normal mammary epithelial cells (184 hterts). As seen in Figure 4.9c, these compounds had no effect on normal cells at 100μM.

4.3 Discussion

We have applied two parallel approaches to find novel off-patent drugs able to inhibit the TNBC drug target RSK. The five compounds we chose to further experimentally test were a mix of HTS-selected (menadione and ellipticine), VS-selected (apigenin), and combination-selected (kaempferol, luteolin) hits. We found that these compounds could block the phosphorylation and nuclear translocation of YB-1, indicative of an inhibitory effect on RSK catalytic activity. In addition, the drugs were able to inhibit the growth of two TNBC cell lines while not affecting normal mammary epithelial cells, indicating a selective ability to kill TNBC breast cancer cells. Apigenin and luteolin were less potent towards MDA-MB231

cells in comparison to SUM149 cells. However, this was a reasonable result as SUM149 cells are driven by EGFR overexpression and signaling through the MAPK pathway [216], whereas MDA-MB-231 also exhibit aberrant PI3K/AKT signaling parallel to the MAPK pathway [255]. Thus a RSK inhibitor, acting downstream of the MAPK pathway, would be less effective in MDA-MB-231 cells.

It is important to consider that many of these off-patent drugs also work through other biological avenues, especially since they were originally designed against other drug targets. This may have accounted for the flavonoids apigenin and luteolin inhibiting TNBC cell growth more strongly than kaempferol despite weaker inhibition constants. Strong growth inhibition of ellipticine could be partly due to its modulation of P53 or perhaps through as-yet unidentified off-target interactions. However, its strong inhibition of RSK activity, P-YB-1 signaling, and TNBC cells suggest that it has potential to be repositioned towards RSK-related cancers and a RSK-related subset of TNBCs.

Several novel TNBC repositioning candidates were discovered in this study. Luteolin and apigenin are promising in that they are bioflavonoids that can be taken through dietary means like parsley or celery. However, it may be difficult to absorb them into the bloodstream at high enough concentrations to inhibit RSK. Though some oral drugs can show plasma concentrations of 100μM, like the MEK inhibitor CI-1040 [256], previous studies have shown that luteolin bioavailability is only about 14nM [257]. Flavonoids can be used as dietary recommendations to accompany therapeutic treatment, to additively inhibit RSK activity; however, the low bioavailability suggests that there may not be enough free drug to significantly inhibit RSK. Though menadione was not validated as a RSK-YB-1 inhibitor during the secondary screen and considered a false positive of the HTS screen, I note that it did show strong inhibition of RSK-S6K peptide. Thus, the strong TNBC growth inhibition of menadione may be due in part to its ability to inhibit other RSK functions aside from YB-1 phosphorylation. The compound ellipticine was especially promising as it is already an anti-cancer agent, based on its ability to rescue mutant *P53* transcription and modulate *P53* nuclear localization [258, 259].

Unfortunately, none of the VS peptide-site predictions showed activity in the HTS screen. However, designing potent small molecule inhibitors of protein-protein interactions still remains an extremely challenging task today. Not only are protein-protein interaction sites shallow and large (and thus generally unable to form a few strong interactions with a small molecule), they are also understood to be difficult to detect in HTS methods [260]. Thus, it will be extremely difficult to develop novel protein-protein inhibiting drugs until better HTS and VS methods are in place.

The positive compounds were substantially different between the HTS and VS lists, with only six compounds were common to both. There are many possible explanations for the low overlap, illustrating the various advantages and limitations of HTS and VS methods. First, docking uses a rigid protein and the correct protein conformation needed to bind the specific drug may not be in RSK structure database. This was apparent for the rigid-planar inhibitor staurosporine, which could only dock well to its cognate structure. Ellipticine and menadione also appear to be this type of compound, with connected rings and no rotatable bonds. Thus the poor amenability of rigid compounds must be taken in consideration when conducting VS. In contrast, HTS methods would not have difficulties detecting such compounds. Second, the docking process lacks water and cofactor molecules. This could explain the less than optimal icm- scores obtained for many of the HTS hits. An alternate explanation for poor scores could be an incomplete scoring function. For instance, ICM was not able to determine why kaempferol was a much stronger binder than the very similar compounds luteolin and apigenin. ICM also predicted the flavonoid epicatechin which did not show any RSK inhibition during HTS. Incomplete scoring functions and/or rigid protein conformations may have impacted the docking of these compounds. To mitigate this problem, other docking software such as Glide [136] or GOLD [135] can be used as secondary virtual screens. Fourth, my docking method was targeted towards only two binding pockets – the ATP and peptide pockets. However, a binding site prediction using PocketFinder [78] shows that there are many potential small molecule binding pockets in the protein (Figure 4.10). HTS screens would be able to detect binding to any of them if the event affected RSK catalytic activity. Lastly, HTS results are also prone to false positive results, as seen with menadione showing poor activity in a secondary *in vitro* kinase assay using the same YB-1 substrate as the HTS.

In addition the steroid hormones may also have been false positive hits since they were previously found to activate p90RSK.

The docking and thresholding criteria developed in this analysis will be useful in future studies. There are some docking studies using ensembles of protein structures [261], some using homology models [262], some using consensus scoring criteria [263] and most using visual criteria (though the latter is usually based on expert judgment rather than strict criteria). Here, I have incorporated all of the above methods to improve the accuracy of my docking pipelines and have delineated criteria that could enrich for known RSK inhibitors by 26 times and HTS inhibitors by 11 times. These criteria would be especially useful for future screens with larger databases, or with luteolin-like compounds. Furthermore, future RSK crystal structures bound to different inhibitors can be added to the structure database to improve the collection of RSK protein conformations.

Overall, the docking methods and thresholds developed could select some of the strongest HTS inhibitors (ellipticine, kaempferol) while only needing to experimentally test the top 2% of the Prestwick library. The docked results also correctly predicted false positive HTS hits. VS methods are thus a powerful complementary approach to HTS methods to improve efficiency of screening, as well as accuracy of discovered hits. Furthermore, knowing the binding mode for a strong inhibitor is especially useful for designing derivatives. However, for both methods, low-throughput follow up assays as well as growth inhibition and signaling assays are essential for assessing the true utility of the top hits.

4.4 Methods

4.4.1 Sequence alignments

Protein sequences of the RSK kinases and NTKDs were downloaded from UniProt [191] and aligned using standard settings in the ICM graphical user interface [87]. Only one representative structure from each publication was chosen for the alignment in Figure 4.1.

4.4.2 Molecular docking using ICM

Drugs were docked to target receptors using the ICM virtual library screening (VLS) module. This method performs rigid-receptor flexible-ligand docking using a two-step Monte Carlo minimization method and energy scoring function to sample ligand conformations and select the best docking hits. MMFF partial charges and ECEPP/3 force-field parameters were used. To ensure a sufficient coverage of the docking energy landscape, I docked every drug-target interaction 10 times. High-throughput docking was performed on a Linux cluster with 175 licenses of ICM. As a given protein may have several structures, each of which with more than one pocket, I docked all pockets to a drug, and the best scoring interaction is selected to be the representative protein-drug score.

I set the minimum icm- and pmf- score threshold to be (-30, -30). On a 2D score plot (such as Figure 4.4), straight lines of negative slope and intercept were used to separate the predicted hits into passing-threshold and failing-threshold groups. The best predictions passing the threshold were in the bottom left corner of the plot. The failing predictions fell in the trapezoid formed by the line, icm-score=-30, pmf-score=-30 and the minimum and maximum icm- and pmf- scores. For each line, I calculated the density of failing-threshold hits in the trapezoid, and chose the line passing 10% of the screened compounds that also had the densest set of failed points.

4.4.3 Inverse docking

The 10 known RSK inhibitors were docked to the database of 252 unique human protein drug targets annotated by DrugBank. It consists of target proteins for which at least one known drug could be docked with a good score, and represents a selection of targets that are more reliable for my docking system. Since then, I have updated it with the latest DrugBank and PDB information. Inverse docking was carried out as the docking in 3.4.4, though each drug-target interaction was docked 5 times instead of 10. Since I could not apply a consensus score threshold for each protein (each protein was only docked to 50 drugs, not allowing enough points for a reliable autonomous threshold), I applied a simple -32 score threshold.

4.4.4 Creating a model of RSK bound to YB-1 peptide

Models were examined with the ICM protein health option [264], which, similar to the pmf-score, compares the energy profile of each amino acid the model to existing PDB structure profiles to reduce statistically unlikely energy strain in the model.

A model of the RSK-YB1 peptide-binding complex was built through an iterative docking procedure, using a known binding complex of AKT-GSK3B peptide (PDB id 1o6k) [265] as a template. This structure was chosen for its structure quality (resolution=1.70Å, R-value=0.205, R-free=0.234, no gaps near peptide) and because the protein was already in a peptide-bound conformation. First, a homology model of RSK was built upon 1o6k, where the sequence identity was 46% over the kinase domain, and 67% within residues 3.5Å to the peptide-binding site. Protein regularization was then performed to optimize the covalent residue geometries. The peptide-binding site was defined to be a large 3-dimensional box containing all residues within 3.5Å of the known GSK3B substrate position.

Initial rigid-receptor docking of a chemically optimized YB-1 peptide structure into the peptide-binding site (receptor) was unsuccessful, due to the vast conformational search space of a flexible peptide. The larger number of rotatable bonds in a peptide compared to a small molecule exponentially increased the number of potential binding conformations to be sampled during docking. Increasing the thoroughness parameter of docking did not improve the results, as molecular docking is a technique generally suited for small molecules and not peptide chains. Including the two manganese ions and small molecule ATP-analog inhibitor in known positions (extracted from the original 1o6k structure) also did not generate any reasonable YB-1 docked binding conformations. Instead, knowing that the YB-1 peptide is phosphorylated at Ser6 (Ser102 in the full protein) [245], I visually scanned through the list of predicted conformations in the results stack and selected those with Ser6 in a position allowing for phosphorylation. For this step, the serine phosphorylation site on the GSK3B peptide was used as a reference. These YB-1 conformations were used as starting points for further docking into the RSK peptide-binding site (instead of re-optimizing the peptide each time as is the normal docking procedure). The docking was iterated, each time selecting a peptide with similar or better docking score and a better binding conformation by visual

judgment. This process ended when the binding poses and energies of the selected peptide conformation stabilized over five iterations. The docking score of the YB-1 peptide in the protein binding site was -55, very strong by ICM standards, and was comparable to other RSK-peptide complex scores.

All of the computational analyses were performed using the Molsoft ICM 3.5-1m software package [87].

4.4.5 Creating a RSK structure database

I obtained RSK crystal structures from the Protein Data Bank (PDB) [224]. I prepared protein structures for docking using Molsoft's ICM software version 3.6-1c [87], removing metal ions, water molecules, solvent ions, and other ligands from the structures. I added hydrogen atoms to the structures then optimized in their positions. To predict potential ligand-binding pockets in the proteins, I used the PocketFinder [78] method in ICM, which calculates a transformation of the van der Waals energy for an aliphatic carbon probe on a grid map. The receptor is defined as the box 3.5Å surrounding the pocket, and the three largest pockets were added to the EGFR binding site collection. In total, there were 16 pockets.

4.4.6 Chemicals

The Prestwick Chemical Library (Prestwick Chemical, Washington, DC) was used as part of the Canadian Chemical Biology Network at the University of British Columbia. Kaempferol, ellipticine, menadione, apigenin and luteolin were purchased from Sigma-Aldrich Chemical (Oakville, ON, Canada) and were dissolved in DMSO to stock concentrations of 100mM. Drugs were then further diluted in cell culture medium as necessary to working concentrations.

4.4.7 RSK1 and RSK2 kinase screens

Kinase profiling services were provided by SignalChem (Richmond, BC) using methods as previously described [245]. Briefly, the RSK kinase assays were performed using a synthetic YB-1 cell permeable peptide (YB-1 CPP) [266] that contains the S102 region as the

substrate. For RSK1, the Prestwick Chemical Library was screened at 10 μM against the YB-1 peptide and results compared to a staurosporine control (a broad-spectrum kinase inhibitor) that has 100% inhibitory activity. Compounds with >20% inhibitory activity were considered to be significant RSK inhibitors. For RSK2, the compounds kaempferol, menadione, ellipticine, apigenin, and luteolin were similarly screened against the YB-1 peptide. Drug treatment concentrations in the RSK2 kinase assay were 0.001, 0.01, 0.1, 1.0, 10, and 100 μM. BI-D1870, a known RSK inhibitor [267], was also examined at these concentrations. For each compound, a graph of log concentration (μM) versus % inhibition of RSK2 activity was generated and IC_{50} values were determined. To confirm inhibition of RSK2 activity, we also used a secondary RSK substrate, S6K, and repeated the kinase assay as described above.

4.4.8 Cell culture

The triple-negative breast cancer cell lines SUM149 (Asterand, Ann Arbor, MI) and MDA-MB-231 (American Tissue Culture Collection, Manassus, VA) were grown as previously described [216] in a 37°C humidified incubator with 5% CO2.

4.4.9 Immunofluorescence and western blotting

SUM149 cells were plated on 8-well multi-chamber slides (40000 cells/well), allowed to adhere for 24 h, then treated with 10 μM of each lead compound for 24 h. Immunofluorescence was conducted as previously described [354] using P-YB-1S^{102} and YB-1 antibodies (Cell Signaling, Danvers, MA) with Alexa-Fluor 488 (Invitrogen) secondary and images were acquired on an Olympus BX61 microscope and analysed using ImageJ (NIH, Bethesda, MD). For western-blotting, cell lysates were collected after 24 h drug treatments and immunoblotting was performed as described previously using P-YB-1S^{102}, YB-1, and αβ-tubulin (Cell Signaling) antibodies.

4.4.10 Monolayer, mammosphere and soft agar growth assays

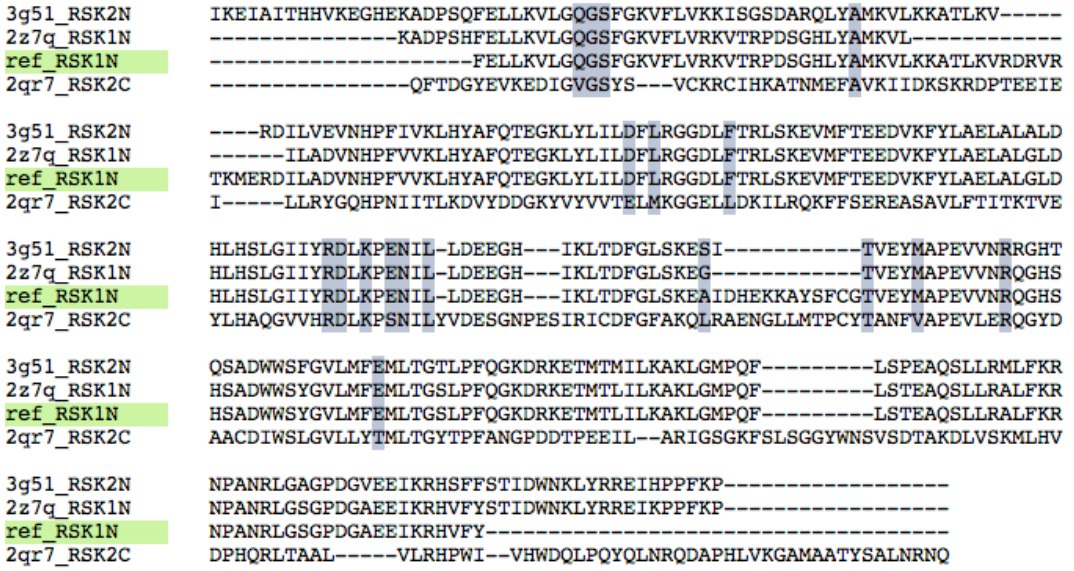
Monolayer growth assays were performed with 5000 (SUM149) or 3000 (MDA-MB-231) cells per well in a 96 well plate. Following 24h after plating, the cells were treated with DMSO, 10 μM or 100 μM of the drugs. The number of cells was counted by high-content

screening as previously described [269] after 72 h drug treatment. Soft agar assays were also performed as previously described [270]. Compounds were added at 10 μM at time of seeding into the top layer and colonies were counted after 28-30 d. Percent change in growth was assessed compared to DMSO control. All growth assays were repeated.

Figure 4.1 Sequence alignments of RSK1 against existing PDB structures.

Comparison of RSK1 N-terminal kinase domain (NTKD) sequence (ref_RSK1N) aligned to a) existing RSK structures and b) an AKT structure. Residues within 3.5Ä of the ATP and substrate binding sites are shaded. Existing RSK1 (2z7q) and RSK2 (3g51) NTKDs have high sequence identity compared to the CTKD (2qr7). However, the NTKDs (3g51, 2z7q) have numerous gaps. b) The AKT structure (1o6k) has fewer gaps when aligned to RSK1, especially near ATP- and substrate- binding site residues (shaded).

a)



b)

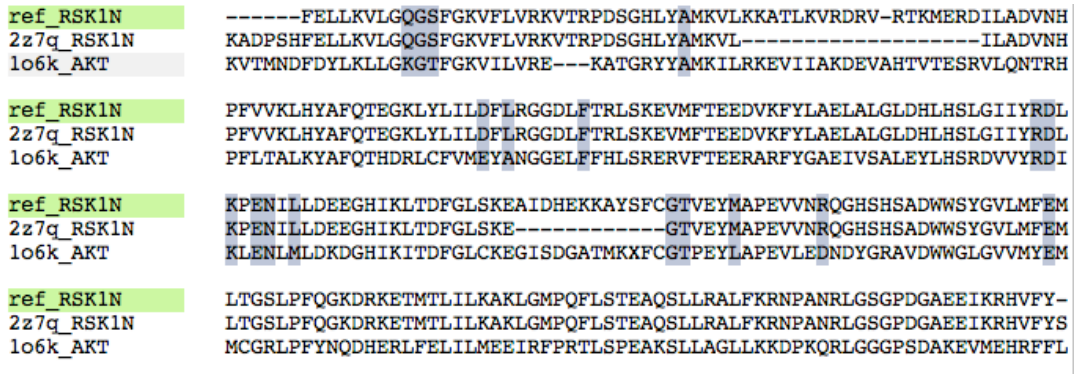
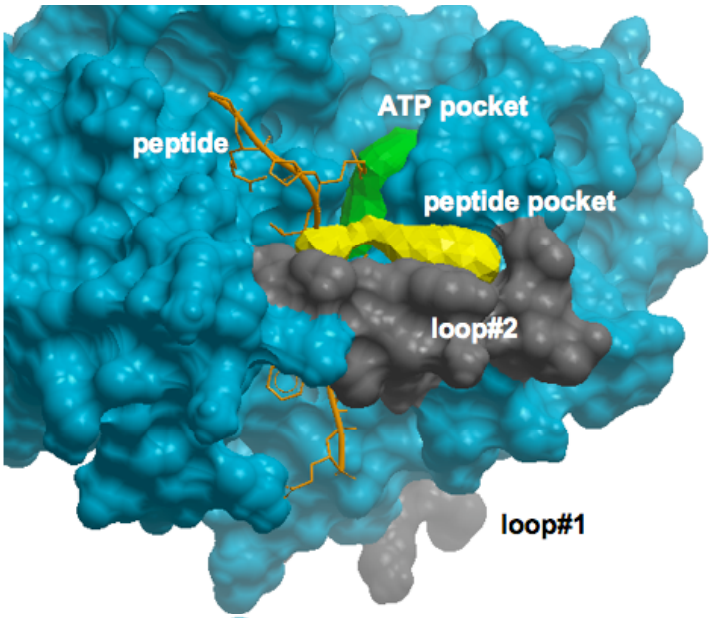


Figure 4.2 RSK homology models.

Two homology models of RSK1 NTKD were built to add extra protein conformations to or structure collection. a) The template was a RSK2 structure bound to an ATP-analog without the aid of metal ions. ICM loop modeling was used to predict energetically favorable positions for the two loops corresponding to gaps in the sequence alignment. b) The template was an AKT structure bound to a substrate peptide. Loops were resolved in this structure (no gaps), thus not requiring further modeling. Both models contained novel peptide-site pockets that were not present in existing RSK structures.

a) Model 1: RSK1 model based on 3g51 (RSK2)



b) Model 2: RSK1 model based on 1o6k (AKT).

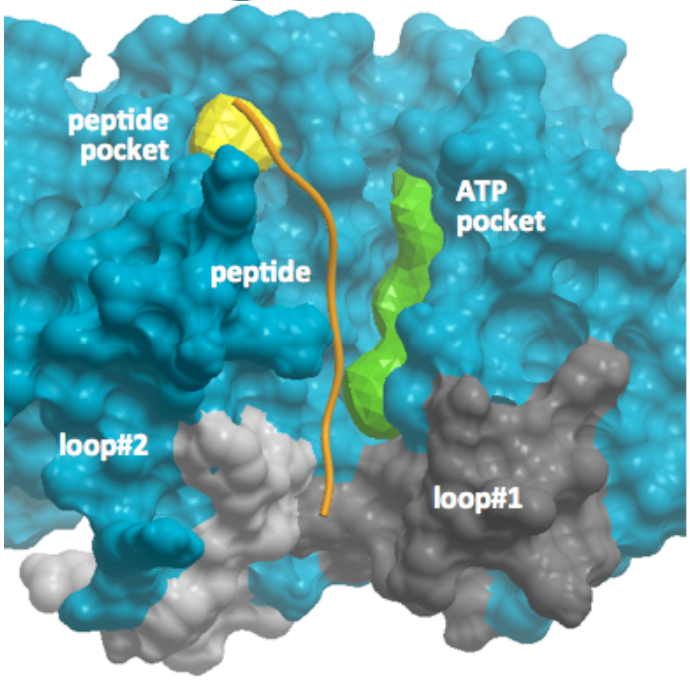


Figure 4.3 A detailed look at Model 2: RSK1 bound to YB-1 peptide.

The YB-1 peptide is shown with ribbon backbone and stick residues. Protein residues involved in hydrogen bonding are shown in black wire format. a) RSK and AKT can both phosphorylate YB-1 and GSK3B. b) Protein surface view of the RSK-YB-1 interaction. c) An alternate view of b) showing hydrogen bonds (green) between the peptide and protein. d) The AKT-GSK3B template complex shows that hydrogen bonds are preserved between corresponding AKT and RSK. Numerous protein residues involved in peptide binding are also preserved between AKT and RSK.

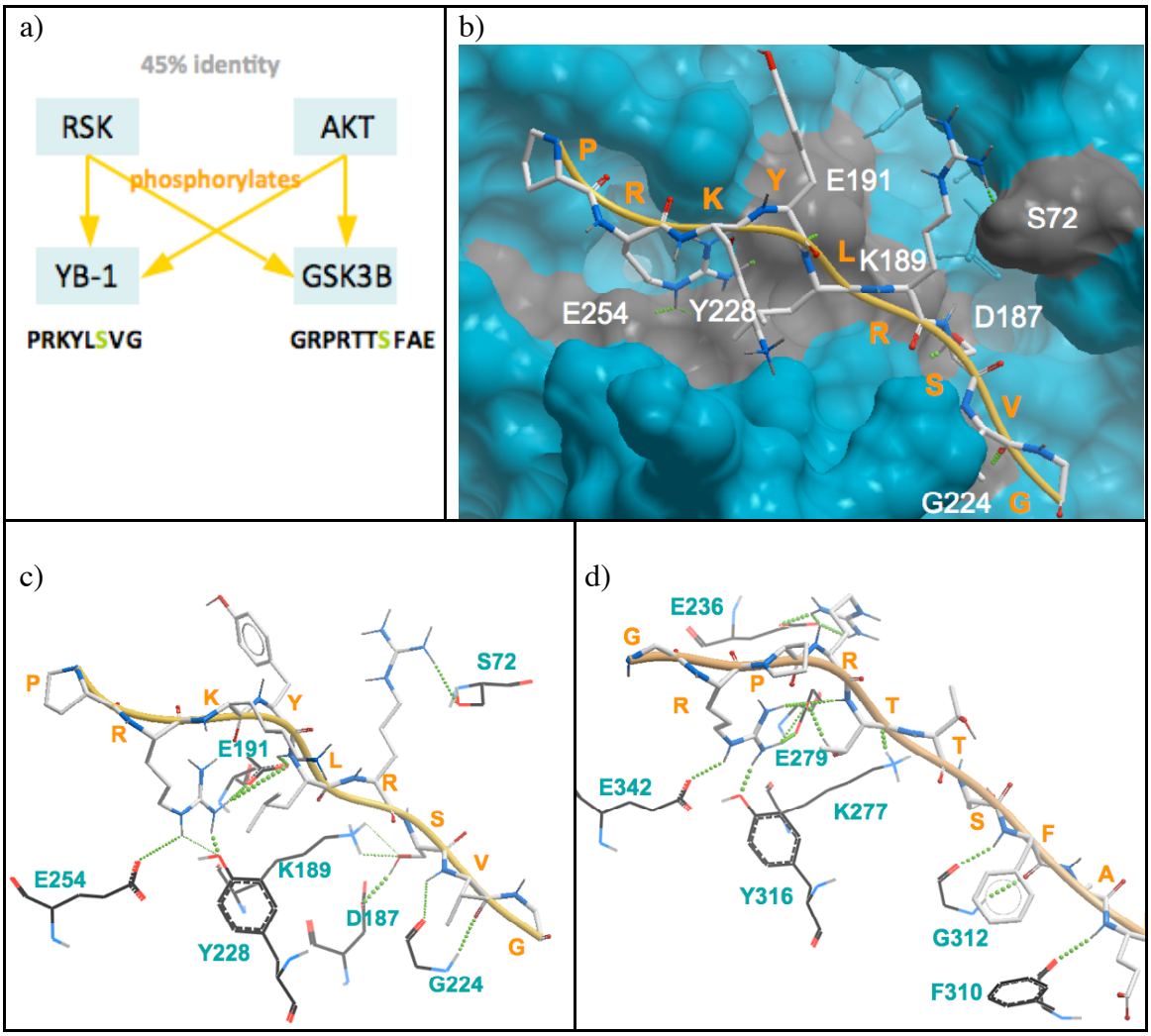


Figure 4.4 Score plot of Prestwick drugs docked to the RSK ATP-binding site.

The (icm-score, pmf-score) pair for each of the 1,120 drugs is plotted in blue. There are 10 extra points, corresponding to known RSK inhibitors collected from the literature (red boxes). A 10% consensus score threshold selected the top 111 score-pairs. These drugs were subsequently docked to a panel of 252 drug targets to determine the protein rank of RSK relative to other targets. Drug compounds that docked to RSK better than at least 232 other targets are shown in black (purple circles, protein rank ≤20). A bare minimum icm-score threshold of -20 was used to remove the worst scoring compounds.

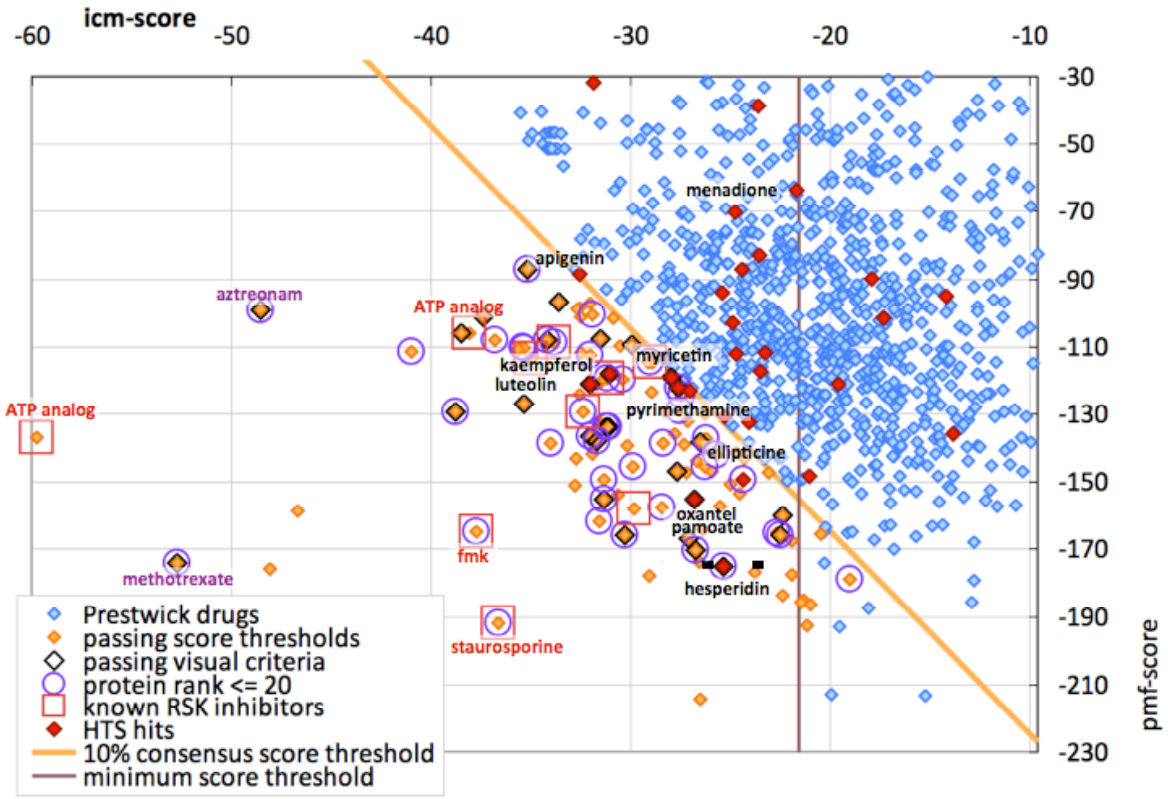


Figure 4.5 Score plot of Prestwick compounds docked to the RSK substrate-binding site.

The (icm-score, pmf-score) pair for each of the 1,120 drugs is plotted in blue. A 10% consensus score threshold and a protein rank ≤20 criterion were used to select the top seven scoring drugs.

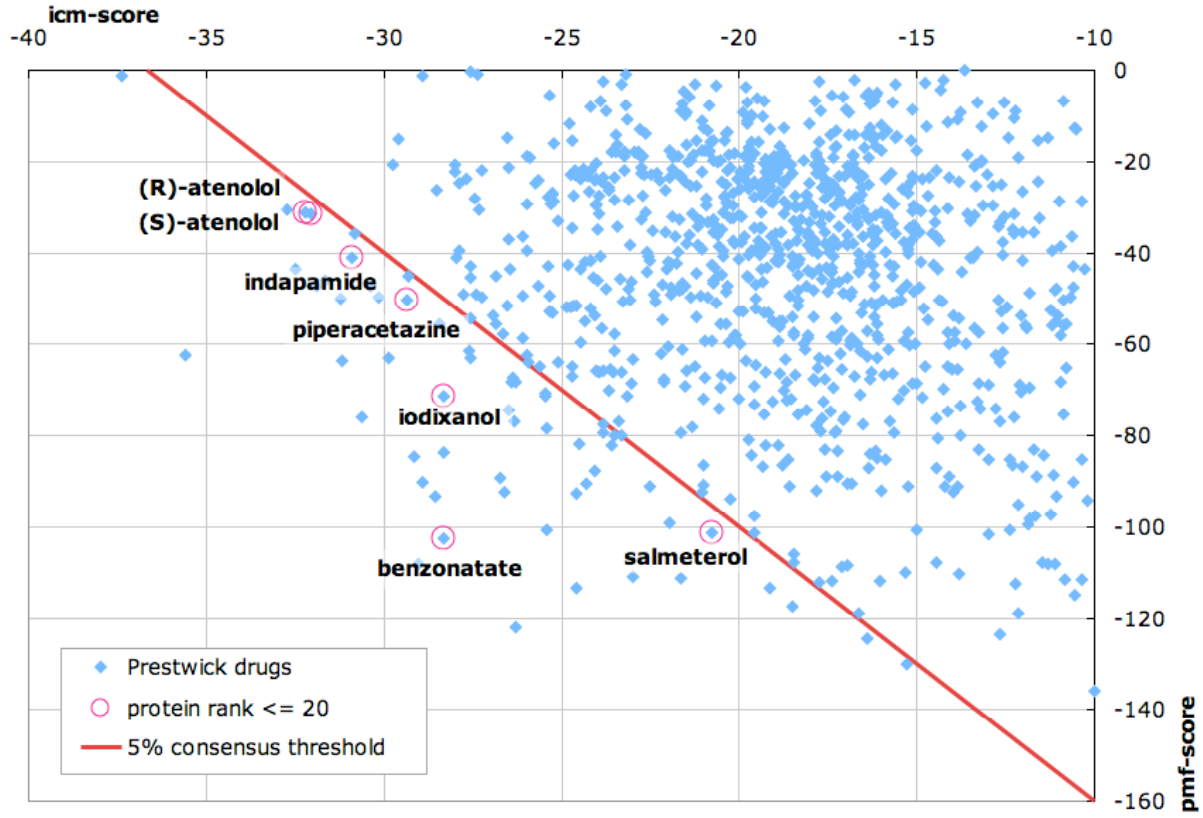


Figure 4.6 Prestwick drugs that modulate the activity of RSK1 in a HTS.

The Prestwick Chemical Library of 1,120 off-patent compounds was screened against RSK1 in an ATP-competitive high-throughput assay, using YB-1 peptide as the substrate. The 32 drugs that inhibited RSK1 activity significantly (>20%) are shown here.

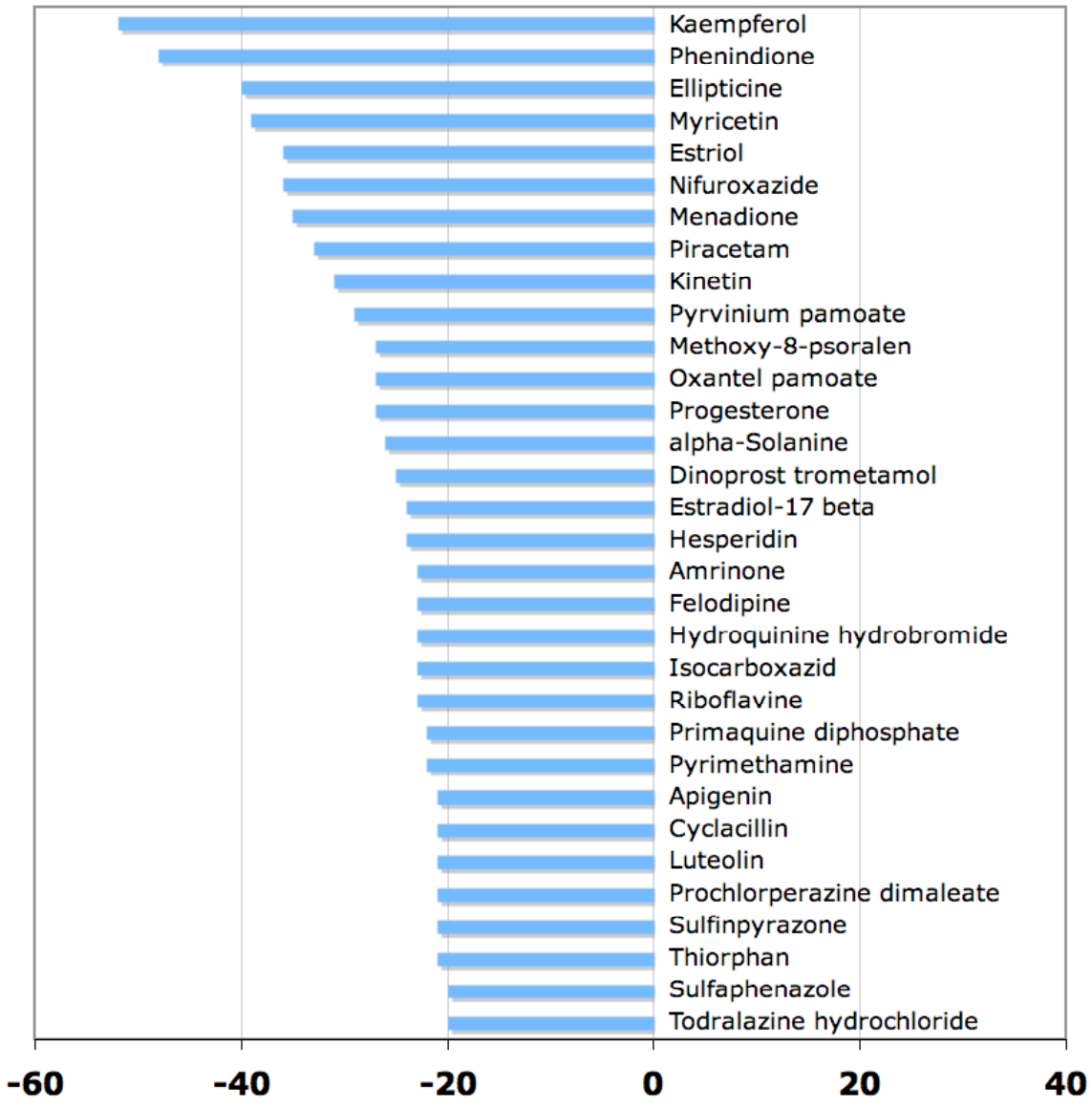
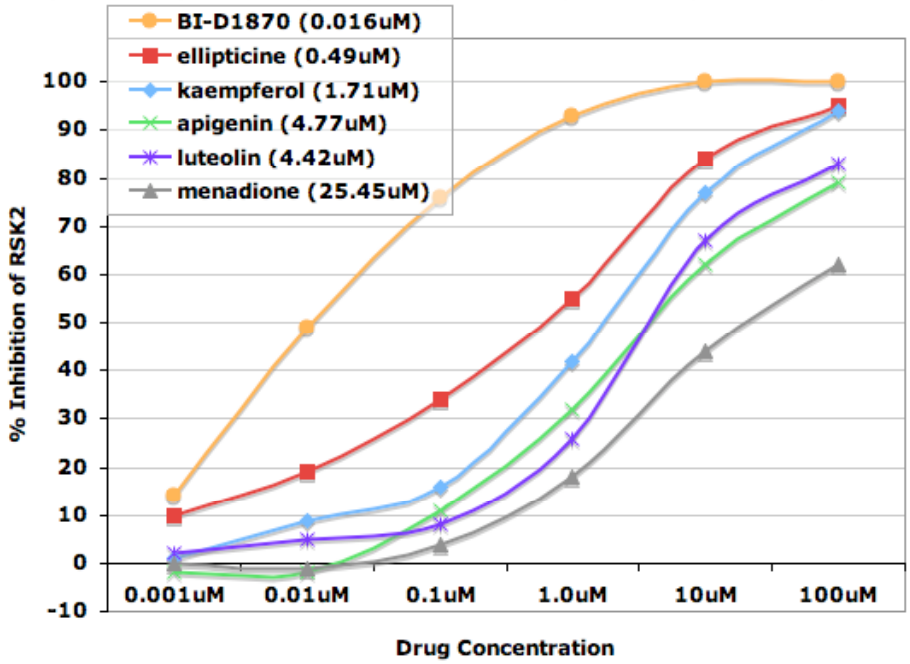


Figure 4.7 Secondary i*n vitro* screens of lead HTS/VS compounds confirm activity.

Five lead compounds from the HTS and VS analyses were analyzed using low-throughput RSK2 ATP-competitive kinase assays, with BI-D1870 as a positive control drug. The IC_{50} value for each compound was determined.

a) Inhibition curves for 6 selected drugs against RSK2 using the YB-1 substrate



b) Inhibition curves for 6 selected drugs against RSK2 using the S6K substrate.

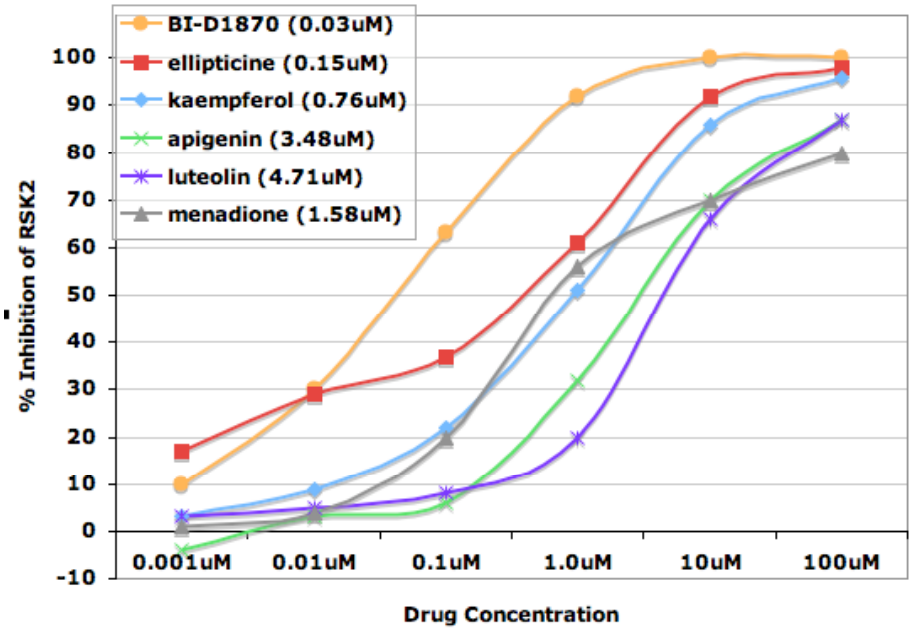
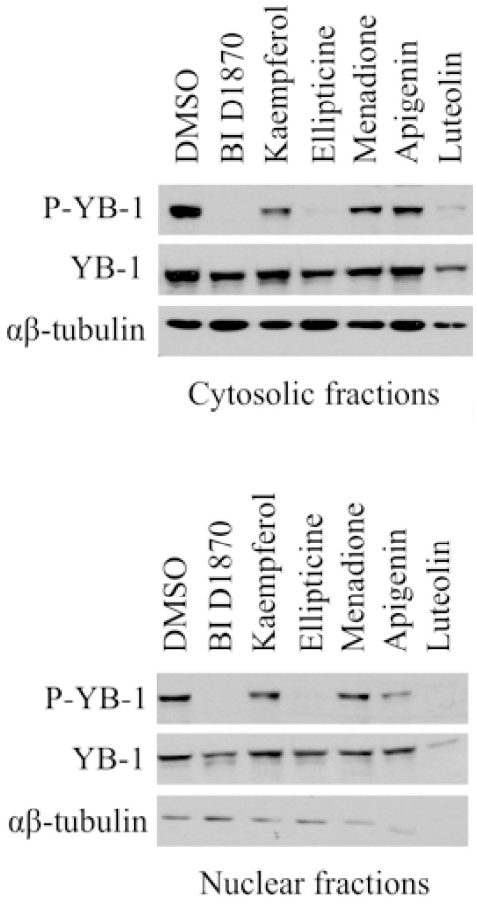


Figure 4.8 Lead compounds block YB-1 activation and nuclear translocation in SUM149 cells.

a) Levels of P-YB-1, a downstream target of RSK, are diminished in SUM149 cells after 24-hour drug treatment. For the control drug BI-D1870 and four of the five lead compounds, there was a marked decrease in P-YB1 levels compared to total YB-1 in both the cytosol and the nucleus. Only menadione had little effect on YB-1 signaling in both fractions.



b) The immunofluorescence of nuclear P-YB1 is markedly reduced in SUM149 cells after 24-hour drug treatment of all lead compounds.

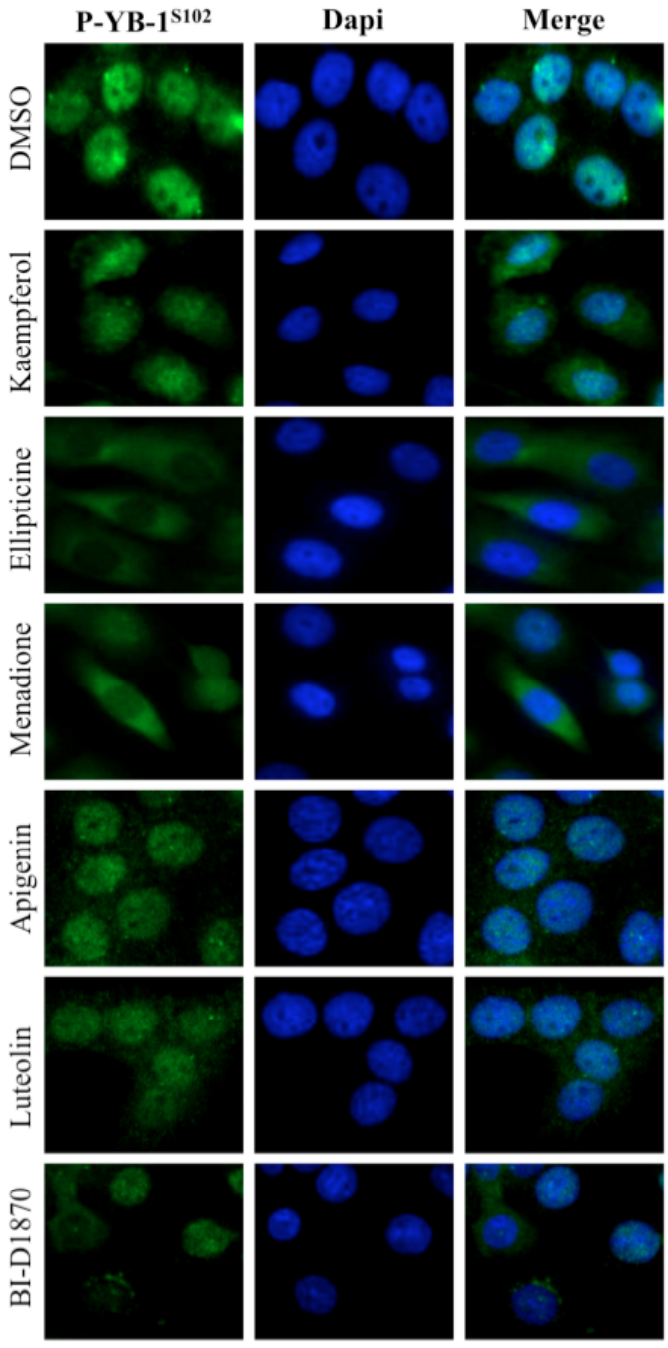
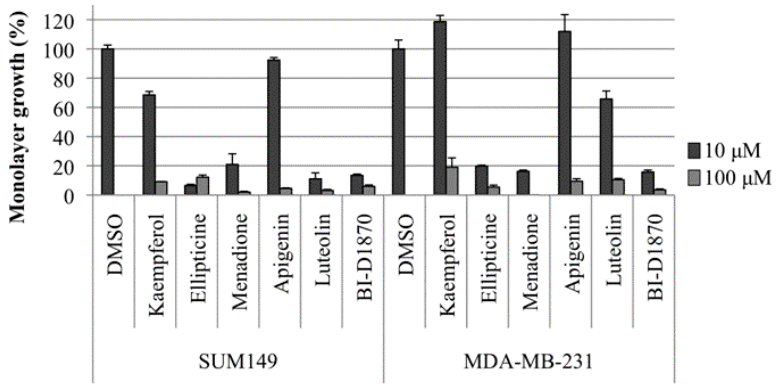


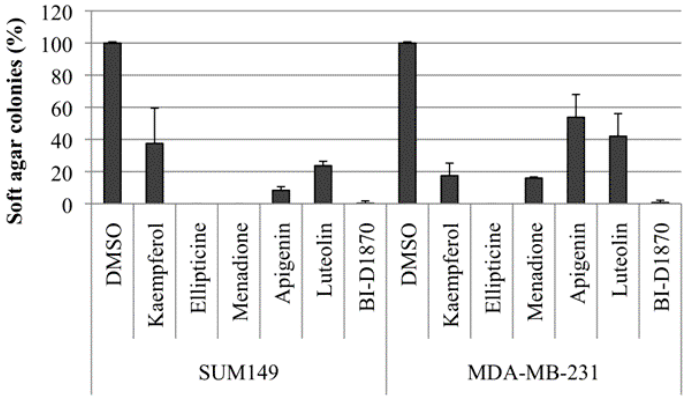
Figure 4.9 Lead compounds inhibit growth in TNBC cell lines.

SUM149 and MDA-MB-231 cells were assayed for growth in monolayer and soft agar after treatment with the five lead compounds. For the TNBC assays, BI-D1870 was used as a positive control and DMSO a vehicle control. a) In the monolayer assay, the cell-lines were treated in triplicate with 10μM or 100μM of drug for 72 hours. b) In the soft agar assay, 10μM of drug was added to the top layer at the time of seeding each cell line. Colonies were counted after 28 days. c) Growth in normal mammary epithelial (184htert) cells was not inhibited by any of the lead compounds at 100μM drug concentration.

a)



b)



c)

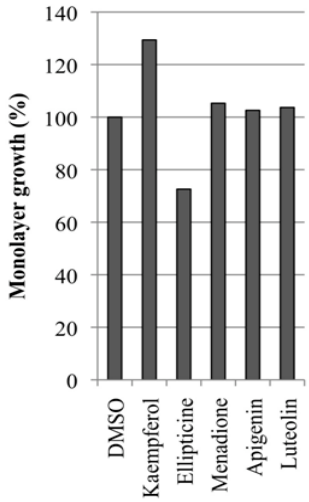


Figure 4.10 Potential small molecule binding pockets in one RSK kinase domain.

Pockets were predicted using the ICM Pocketfinder method. It can be seen that the RSK NTKD has many deep cavities that are compatible for small molecule ligand binding.

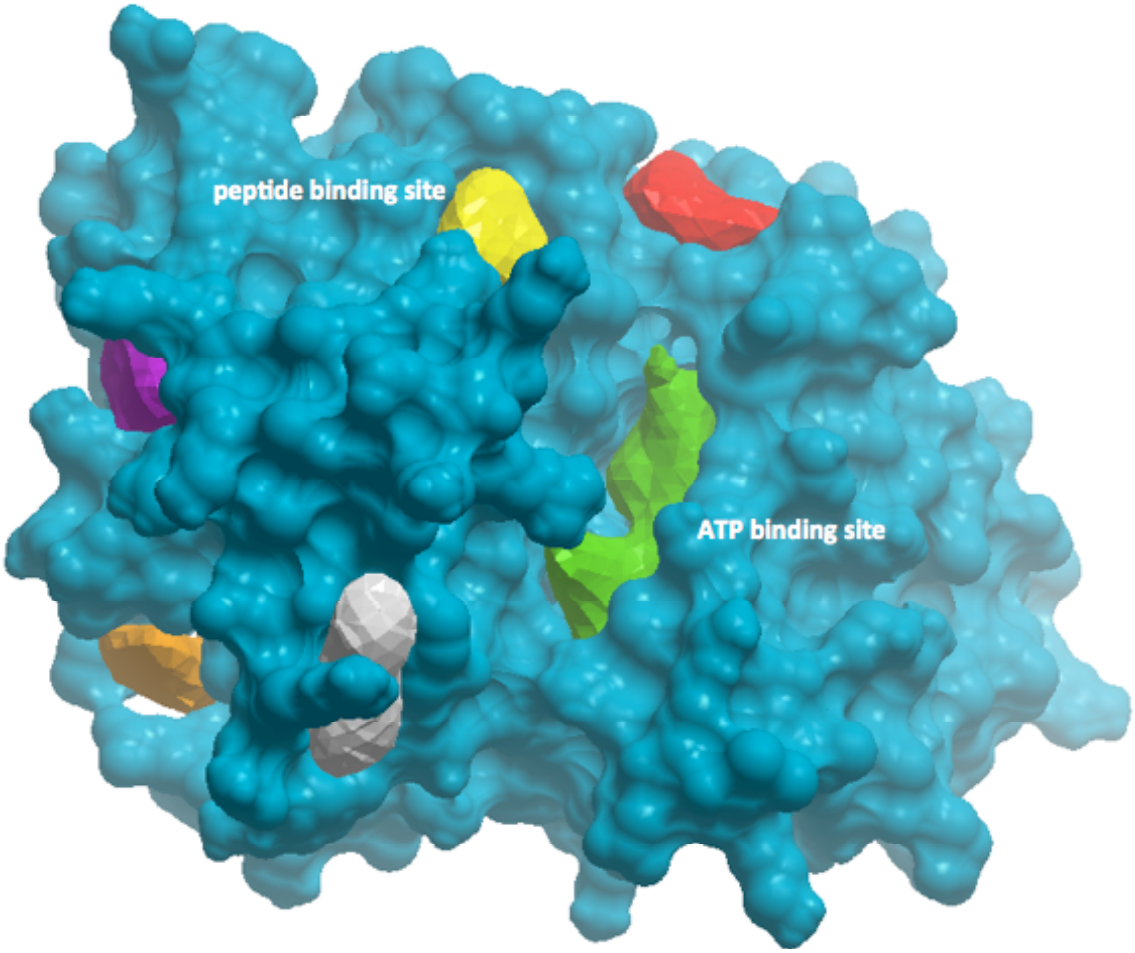


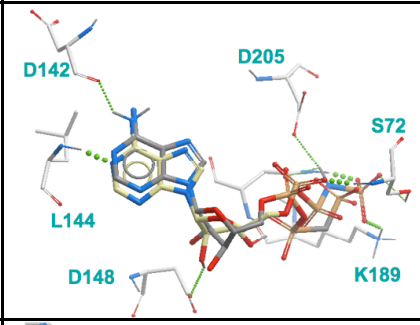
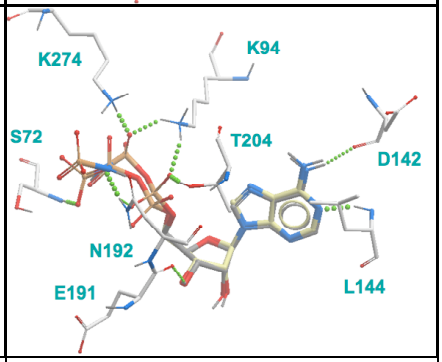
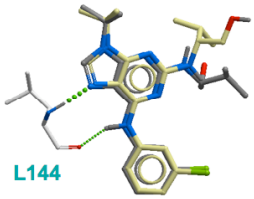
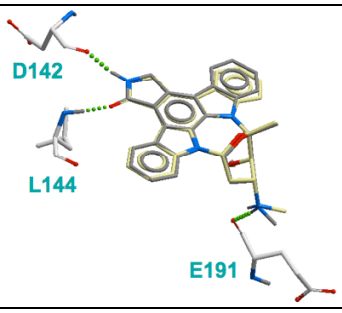
Table 4.1 Existing crystal structures of the RSK protein

Existing RSK protein 3D structures represented in the binding-site database. For each structure, details about the isoform, N-terminal or C-terminal location, and active site ligand, are provided.

PDB ID	Resolution (Å)	RSK isoform	Kinase domain	Active site ligand	Reference
2wnt	2.4	RSK1	N-terminal	-	[76]
2z7q	2.0	RSK1	N-terminal	ATP analog AMP-PCP	[271]
2z7r	2.0	RSK1	N-terminal	staurosporine	
2z7s	2.1	RSK1	N-terminal	purvalanol A	
3g51	1.8	RSK2	N-terminal	ATP analog AMP-PNP	[272]
2qr7	2.0	RSK2	C-terminal	-	[273]
2qr8	2.0	RSK2	C-terminal	-	

Table 4.2 Docking known compounds to RSK structures as positive controls.

The docked compound conformations are shown in yellow and protein residues forming hydrogen bond contacts with the compounds are shown in white. Since each compound was docked to multiple RSK structures, the isoform of the best scoring RSK protein is listed. The docked conformations of cognate scenarios a-d) were compared to their bound conformations (grey) in existing PDB complexes both using RMSD.

	Inhibitor	RSK isoform	icm-score	pmf-score	RMSD (Å)	Docked conformation
a)	AMP-PCP	RSK1	0.57	-111	2.16	
b)	AMP-PNP	RSK2	-54	-138	1.30	
c)	purvalanol A	RSK1	-31	-133	1.93	
d)	stauro-sporine	RSK1	-37	-190	0.31	

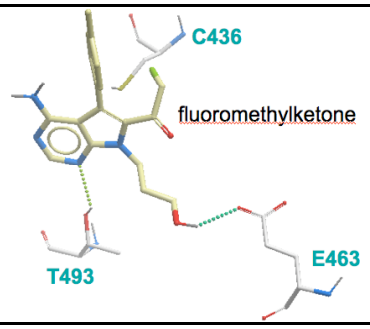
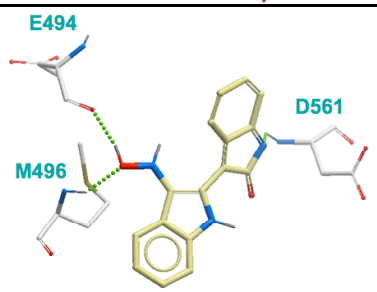
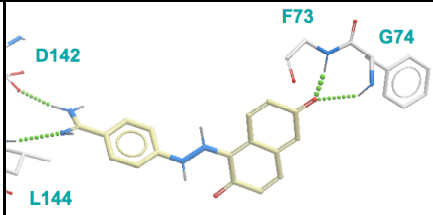
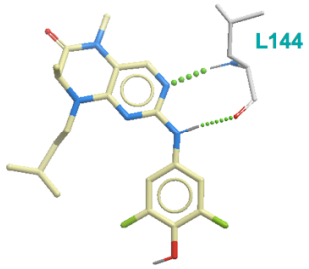
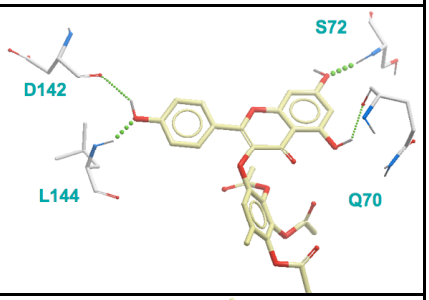
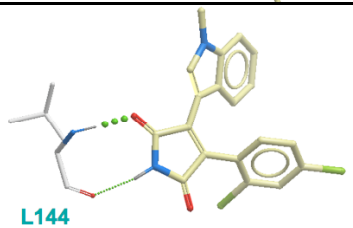
	Inhibitor (IC_{50})	RSK isoform	icm-score	pmf-score	RMSD (Å)	Docked conformation
e)	fmk (15nM)	RSK2	-36	-169	-	
f)	indirubin-3-oxime (0.1μM)	RSK2	-34	-133	-	
g)	nsc 356821 (1μM)	RSK1	-35	-113	-	
h)	BI-D1870 (15nM)	RSK1	-29	-167	-	
i)	sl0101 (90nM)	RSK1	-29	-138	-	
j)	SB 216763 (0.1μM)	RSK1	-28	-133	-	

Table 4.3 Comparison of the effects of using multiple crystal structures.

Below are the distributions of icm-scores for each compound across the different RSK structures. Interactions with good docking scores are shaded in orange, while the best score for each compound is shaded in blue. Overall, there is a wide spread of orange and blue, showing that each structure binds well to a different set of known inhibitors. 2z7q could not dock well to any known RSK binders and was removed from my database. In addition, the docking method was able to discriminate between N-terminal and C-terminal binders, since NSC356821, sl0101, and BI-D1870 are known to be NTKD inhibitors while fmk is a known CTKD inhibitor.

Several compounds are abbreviated with PDB chemical shorthand. STU: staurosporine. p01: purvalanol A. IRB: indirubin-3-oxime.

		2z7q	2z7r	2z7s	model1	model2	3g51	2wnt	2qr7	2qr8
		RSK1N					**RSK2N**	**RSK1C**	**RSK2C**	
a)	AMP-PCP	0.6	-28.3	-20.0	-22.0	-31.6	-38.5	-22.4	-29.6	-23.6
b)	STU	0.6	-35.8	-24.4	-16.4	1.1	-17.4	-8.5	-22	-18.0
c)	P01	-22.3	-31.6	-30.7	-28.3	-5.4	-29.6	-26.6	-16.8	-18.7
d)	AMP-PNP	-14.4	-28	-14	-36.7	-41.5	-60.2	-6.9	-29.6	-4.1
e)	BI-D1870	0.2	-26.2	-29.4	-25.2	-19.4	-27.7	-12.7	-19.9	-17.2
f)	fmk	-10.3	-34.2	-20.9	-17.2	-12.3	-23.7	-27.3	-35.6	-38.1
g)	SB216763	-19.6	-28.7	-27.6	-9.8	-19	-16.8	-29.6	-32.8	-26.7
h)	IRB*	-22.3	-29.1	-23.2	-10.1	-21.4	-17.7	-20.8	-33.8	-30.0
i)	nsc356821	-10.3	-24.5	-27.8	-29.7	-35.4	-28.3	-28.2	-32.3	-29.6
j)	sl0101*	-5.6	-19.9	-29.2	-29.2	-13.2	-29.5	-20	-17.7	-19.1

Table 4.4 The top 21 ranking inhibitors of known RSK inhibitors.

The top predicted off-target interactions for the ten RSK inhibitors, passing stringent criteria: 10% consensus score threshold as well as drug rank ≤ 15, protein rank ≤ 10, pmf-score rank ≤ 20, and pmf-score protein rank ≤ 25. Finally, 21 interactions out of a potential 10x252 = 2520 were selected. Many of the top protein-drug interactions predicted by docking scores and ranks are also true binders or likely true binders based on the literature.

Abbreviations: prk: protein rank. drk: drug rank. STO: staurosporine. IRB: indirubin-3-oxime. p01: purvalanol A.

Known inhibitor	Protein	icm-score	pmf-score	icm drk	icm prk	pmf drk	pmf prk	True binder?	Reference
AMP-PCP	AK1	-40.0	-187	1	9	1	2	likely binder	These ATP analogs are routinely used as kinase inhibitors in crystal structure determination.
AMP-PCP	UCK2	-43.3	-170	2	6	2	6		
AMP-PNP	EGFR	-31.2	-186	5	7	2	14		
AMP-PNP	GSK3B	-33.1	-197	9	6	1	9		
AMP-PNP	PDPK1	-33.9	-212	2	5	1	4		
BI-D1870	RARA	-32.2	-173	2	1	2	3	-	1
fmk	LCK	-39.4	-177	4	2	3	3	yes	[274]
fmk	SRC	-35.9	-152	13	4	5	22	yes	
fmk	EGFR	-31.4	-157	4	7	5	18	likely	Many strong pyrrolo-pyrimidine inhibitors of EGFR exist (PKI-166, AEE-788) [275]
IRB	NT5M	-34.4	-121	5	7	11	19	-	
IRB	KIT	-35.3	-141	1	6	18	9	-	
nsc356821	RARA	-46.8	-138	1	1	18	9	-	
p01	NR1H2	-32.1	-161	2	4	9	6	-	
p01	ALB	-37.6	-153	7	1	6	9	-	
SB216763	GSK3B	-33.1	-148	10	5	13	14	yes	[274]
SB216763	MAPK14	-32.9	-142	14	6	11	20	weak	
sl0101	SRC	-38.5	-156	9	7	3	16	no	[274]
STO	HCK	-33.9	-198	4	4	1	7	yes	[42]
STO	SRC	-34.1	-189	15	3	1	13		
STO	SYK	-42.5	-215	1	1	1	2		
STO	LCK	-42.0	-192	2	2	1	11		

Table 4.5 Examples of drugs eliminated by or visual criteria.

The docked compound conformations are shown in yellow while protein residues forming hydrogen bond contacts with the compounds are shown in white. These compounds had good icm- and pmf- scores but did not pass visual inspection criteria.

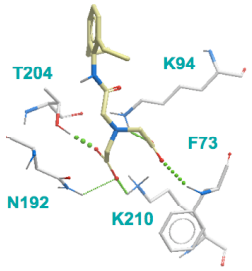
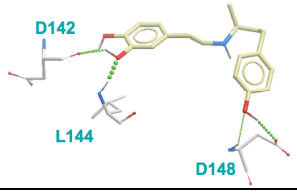
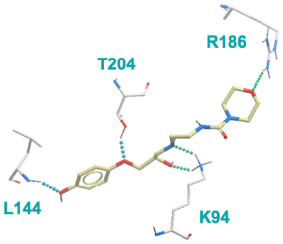
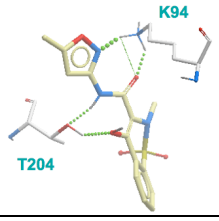
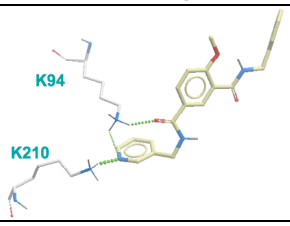
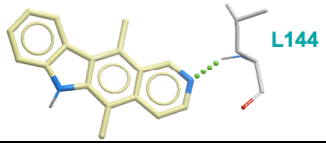
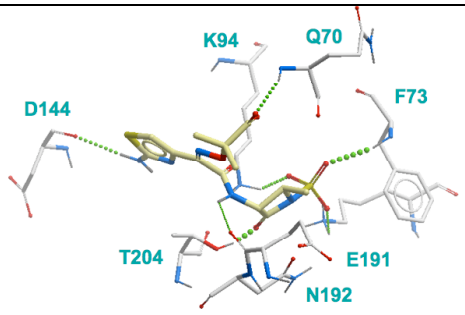
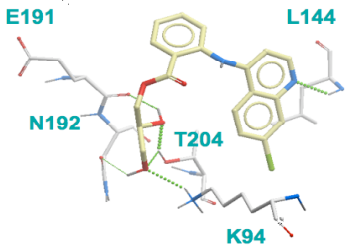
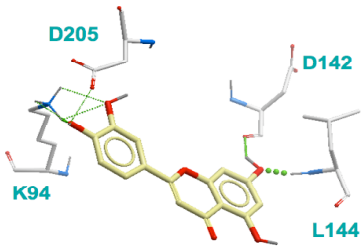
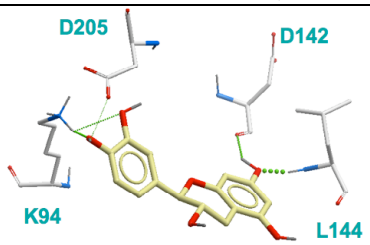
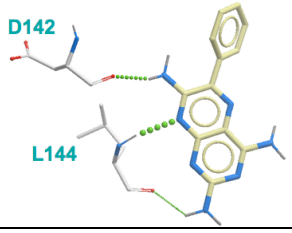
	Compound	Predicted binding conformation	icm-score	pmf-score	protein rank
a)	etifenin		-41	-111	16
b)	dobutamine		-36	-110	4
c)	xamoterol		-30	-145	10
d)	isoxicam		-29	-139	12
e)	picotamide		-27	-139	37
f)	ellipticine		-25	-149	18

Table 4.6 Top Prestwick drugs predicted to bind to the RSK ATP site.

The docked compound conformations are shown in yellow while protein residues forming hydrogen bond contacts with the compounds are shown in white. These compounds were selected through consensus scoring, protein rank, and visual inspection criteria.

Compound	Predicted binding conformation	icm-score	pmf-score	protein rank
aztreonam		-49	-99	11
glafenine		-39	-129	29
luteolin		-32	-121	35
catechin epicatechin		-31.6 -31.5	-134 -134	9 10
triamterine		-31	-37	7

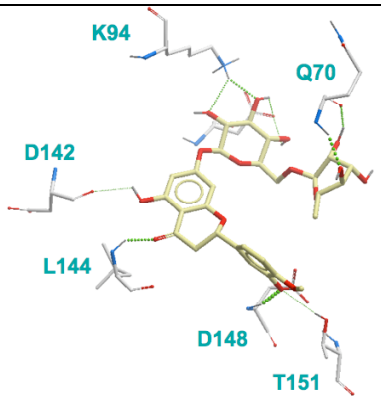
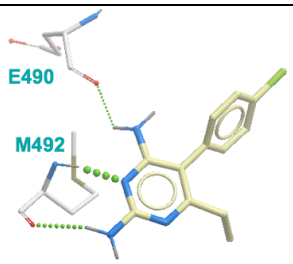
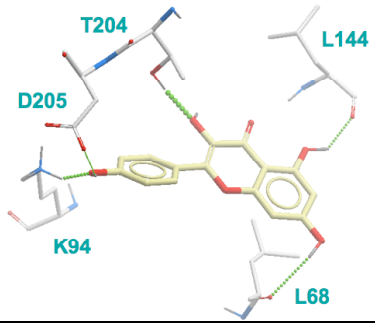
Compound	Predicted binding conformation	icm-score	pmf-score	protein rank
hesperidin		-26	-175	2
pyrimethamine		-28	-122	10
kaempferol		-27	-125	36

Table 4.7 Top Prestwick drugs predicted to bind to the RSK1 substrate-binding site.

These compounds were selected through consensus scoring, protein rank, and visual inspection criteria. Rows are listed by icm-score. On the left, the YB-1 peptide is shown as a ribbon (orange) bound to the RSK surface (blue), and residues that hydrogen bond with the compound are shown in grey. On the right, the compound (yellow) and hydrogen bond forming RSK residues (white) are shown in stick format.

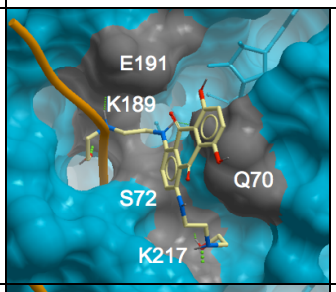
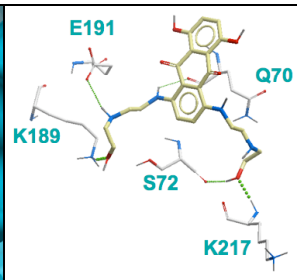
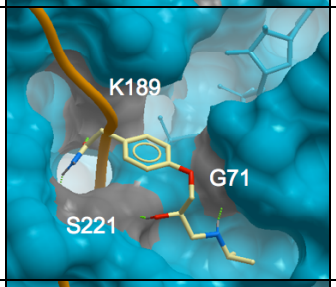
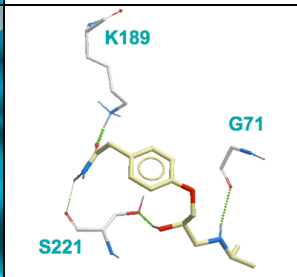
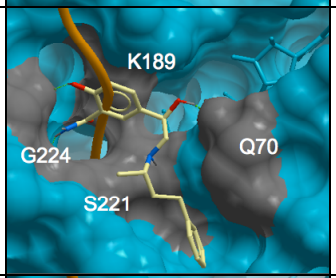
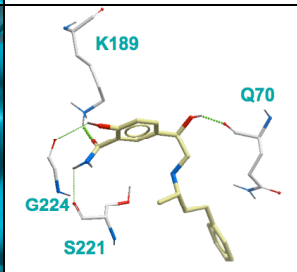
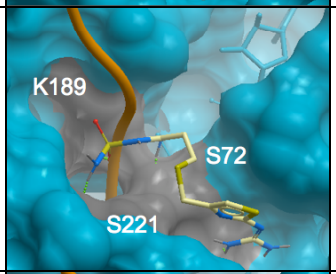
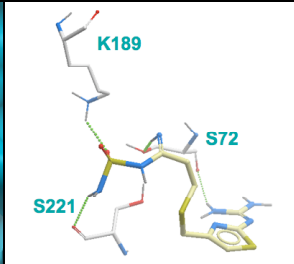
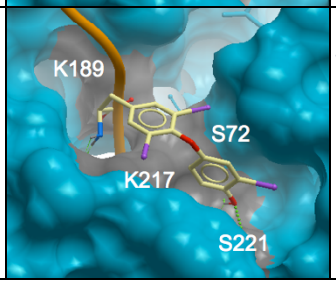
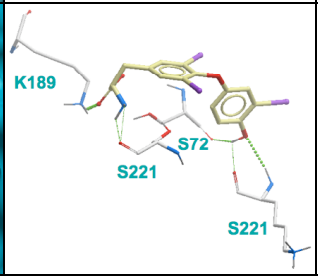
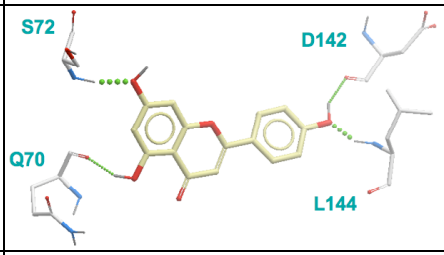
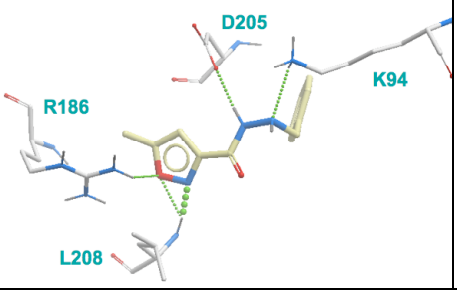
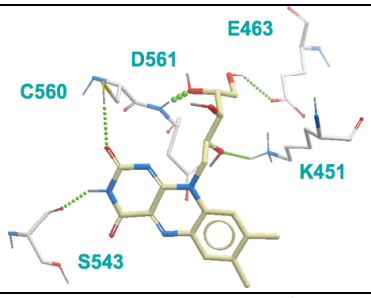
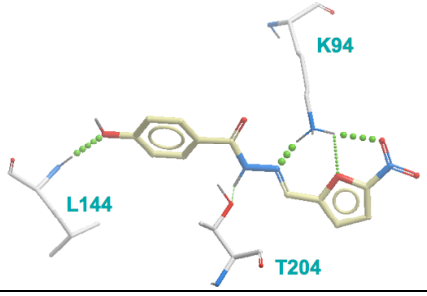
Compound	Predicted binding conformation		icm-score	pmf-score
mitoxantrone			-32.7	-30
(R)-(+)-atenolol (S)-(-)-atenolol			-32.2 -32.1	-31 -31
labetalol			-31.9	-47
famotidine			-31.7	-46
liothyronine			-30.8	-35.8

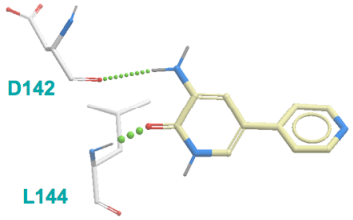
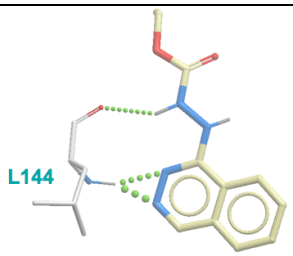
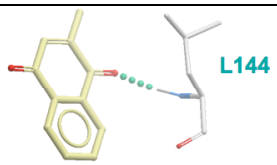
Table 4.8 The enrichment of HTS hits and known ATP binders using traditional ICM scoring cutoffs or the consensus scoring threshold.

	Total # hits	# HTS hits passing threshold	HTS enrichment factor	# known drugs passing threshold	Known drug enrichment factor
No threshold (random)	1130	32	1.0	10	1
10% consensus score	111	7	2.2	10	10.2
10% consensus score & protein rank ≤20	40	3	2.6	6	17.0
10% consensus score & protein rank ≤10	26	2	2.7	6	26.1
10% consensus score & visual criteria	29	6	7.3	-	-
10% consensus score & visual criteria & protein rank <=76 & pmf-score protein rank <=50	19	6	11.2	-	-

Table 4.9 Predicted binding conformations of several Prestwick drugs that had high RSK inhibition in the HTS experiment but did not pass score cut-offs.

The docked compound conformations are shown in yellow while protein residues forming hydrogen bond contacts with the compounds are shown in white. These compounds showed reasonable docked conformations and scores. However, the icm- and pmf- scores were not high enough to pass the consensus score.

	Compound	Predicted binding conformation	icm-score	pmf-score	protein rank	RSK isoform
a)	apigenin		-33	-88	6	RSK2 N-terminal
b)	isocarboxacid		-32	-32	2	RSK1 N-terminal
c)	riboflavine		-27	-123	15	RSK2 C-terminal
d)	nifuroxazide		-26	-131	39	RSK1 N-terminal

	Compound	Predicted binding conformation	icm-score	pmf-score	protein rank	RSK isoform
e)	amrinone		-26	-94	63	RSK1 N-terminal
f)	todralazine		-25	-103	22	RSK1 N-terminal
g)	menadione		-25	-70	19	RSK2 N-terminal

5 Evolution of an Adenocarcinoma in Response to Selection by Targeted Kinase Inhibitors

5.1 Introduction

5.1.1 Challenges in cancer drug discovery

Cancer is a disease arising from uncontrollable cell growth, and is a leading cause of death in the world [276]. Discovering novel cancer drugs is particularly difficult as evidenced by far lower clinical success rates compared to other diseases such as cardiovascular or CNS disorders [13]. Early cancer drugs were cytotoxic and affected all rapidly dividing cells by targeting essential cellular functions such as DNA metabolism, replication, chromosomal segregation, and cytokenesis [277]. A few of these remain as standard chemotherapies today (i.e. methotrexate, doxorubicin), but are not ideal due to their inability to distinguish normal cells from cancer cells. In recent years, targeted therapies aim to inhibit proteins that are involved in the tumor disease mechanism but are less essential in normal cells [278]. However, there must be other factors underlying the low clinical success rates of cancer drug discovery.

It is well known that the same therapy can have different efficacies and toxicities across patient populations. This is due in part to the genetic makeup of the patient, including polymorphisms in drug-metabolizing enzymes like cytochrome P450 [279] or target proteins like EGFR [280]. Another major factor is that cancers are classified by their location of origin and histology, but each class of cancer is actually a collection of diseases with different molecular features. Breast cancer is a prominent example of this phenomenon, having been classified into four broad subtypes and many more sub-subtypes through gene expression profiling [281]. The four subtypes have been shown to respond differently to preoperative chemotherapy [282]. In particular, the triple-negative subtype that does not express estrogen receptor, progesterone receptor, or the HER2 gene is associated with aggressive tumors and poor prognosis. Kidney and ovarian cancer, among many other cancers, are also known to be collections of different diseases at the molecular level with significant variation between patients [283, 284].

To date, gene expression profiling platforms have been constructed for diagnosing the cancer risk or prognosis for individuals, such as MammaPrint or OncoType DX for breast cancer [285]. Fluorescence in situ hybridization (FISH) analysis has been used in the UroVysion kit to detect specific cytogenetic abnormalities in bladder cancer from urine samples [286]. A much higher resolution analysis of genomic aberrations can be conducted through whole genome sequencing of a tumor and its matched normal sample, detecting sequence, copy number, and expression changes, among other aberrations.

5.1.2 Whole genome sequencing for cancer drug discovery

Large-scale sequence analysis of cancer transcriptomes, using expressed sequence tags (ESTs) [287] or serial analysis of gene expression (SAGE) [288, 289], has been used to identify genetic lesions that accrue during oncogenesis. Other studies have involved large-scale PCR amplification of exons and subsequent DNA sequence analysis of the amplicons to survey the mutational status of protein kinases in cancer samples [290], 623 'cancer genes' in lung adenocarcinomas [291], 601 genes in glioblastomas [292], all annotated coding sequences in breast, colorectal [293, 294] and pancreatic tumors [295], searching for somatic mutations that drive oncogenesis.

The development of massively parallel sequencing technologies has provided an unprecedented opportunity to rapidly and efficiently sequence human genomes [296]. The technology has been used to sequence cancer cell line transcriptomes [297-299] and to identify genomic rearrangements in a breast cancer genome [300]. The first human cancer genome was sequenced in 2008 from a patient with Acute Myeloid Leukemia, discovering two genes with mutations already thought to be involved in tumor progression and eight genes with novel mutations [301]. However, methodological approaches for integrated analysis of cancer genome and transcriptome sequences have not been reported, nor has there been evidence presented in the literature that such analysis has the potential to inform the choice of cancer treatment options. We present for the first time such evidence here. The ability to comprehensively genetically characterize tumors at an individual patient level

represents a logical route for informed clinical decision making and increased understanding of these diseases.

5.1.3 Patient history and treatment overview

In this case the patient was a 78 year old, fit and active Caucasian man. A timeline summarizing the patient's examinations and treatments is shown in Figure 5.1. He presented in August 2007 with throat discomfort and was found to have a 2 cm mass at the left base of the tongue. He had minimal comorbidities and no obvious risk factors for an oropharyngeal malignancy. A positron emission tomography-computed tomography (PET-CT) scan identified suspicious uptake in the primary mass and two local lymph nodes. A small biopsy of the tongue lesion revealed a papillary adenocarcinoma, potentially originating from a minor salivary gland. In November 2007 the patient had a laser resection of the tumor and lymph node dissection, with negative final surgical margins. The pathology described a 1.5 cm poorly differentiated adenocarcinoma with micropapillary and mucinous features. Three of 21 neck nodes (from levels 1 to 5) indicated the presence of metastatic adenocarcinoma. Adenocarcinomas of the tongue are rare and represent the minority (20 to 25%) of the salivary gland tumors affecting the tongue [302-304].

The patient then received 60 Gy of adjuvant radiation therapy completed in February 2008. Four months later, although the patient remained asymptomatic, a routine follow up PET-CT scan identified numerous novel small (largest 1.2 cm) bilateral pulmonary metastases but no evidence of local recurrence. There were no standard chemotherapy treatment options for this rare tumor type, so a pathology review was conducted. The review indicated +2 EGFR expression (Zymed assay) and a 6-week trial of the EGFR inhibitor erlotinib was initiated. However, all the pulmonary nodules grew while on this drug and the largest lesion increased in size from 1.5cm to 2.1cm from June 19th to August 18th. Chemotherapy was terminated on August 20th and a repeat CT on October 1st showed growth in all of the lung metastases. At this point, the patient provided explicit consent to pursue a genomic and transcriptome analysis.

5.2 Results

5.2.1 Initial tumor

The patient had multiple pulmonary nodules, and elected to undergo a fresh tumor tissue needle biopsy of a 1.7 cm left upper lobe lung lesion. This was done under CT guidance and multiple aspirates were obtained for analysis.

5.2.1.1 Genome and transcriptome sequencing

Table 5.1 summarizes the patient's tumor and normal DNA and RNA samples that were sequenced and aligned to the reference human genome (HG18). There were 2,584,553,684 tumor DNA reads and 342,019,291 normal DNA purified from peripheral blood cells. This was approximately 18X diploid coverage of the tumor DNA. Whole transcriptome shotgun sequencing (WTSS) [299, 305] was conducted to profile the expression of tumor transcripts. There were 498,229,009 tumor RNA reads, and 62,517,972 normal RNA reads from the leukocytes.

5.2.1.2 Mutation detection and analysis

Our initial analysis of sequence alignments identified 84 DNA putative sequence changes corresponding to non-synonymous changes in protein coding regions present only in the tumor (Table 5.2). Sanger sequencing subsequently validated four changes to be somatic tumor mutations. The vast majority of false positives were due to undetected heterozygous alleles in the germline. Somatic mutations were observed in two well-characterized tumor suppressor genes TP53 (D259Y) and RB1 (L234*). TP53 was within a region of heterozygous loss (LOH) and the truncating mutation of RB1 removed 75% of its coding sequence.

5.2.1.3 Copy number analysis

We concentrated on finding genetic changes likely to affect cellular function, such as changes in gene copy number or protein sequence. Due to our inability to usefully interpret alterations in non-coding regions, such changes were not considered. We compared the relative frequency of sequence alignment derived from the tumor and normal DNA and identified 7,629 genes in chromosomally amplified regions, including 17 genes classified as

being highly amplified. Our analysis also revealed large regions of chromosomal loss (Figure 5.2). Intriguingly, we observed loss of approximately 57 megabases across segments from 18q, on which frequent loss has been observed in colorectal metastases. Other large chromosomal losses were observed in the tumor (17p, 22q and 12p) but did not correlate with previous studies of salivary gland tumors [306-309].

5.2.1.4 Transcriptome analysis

In the absence of an equivalent normal tissue for comparison, we compared expression changes to the patient's leukocytes and a compendium of 50 tumor-derived WTSS datasets (Appendix C). We expected that the compendium would prevent spurious observations due to technical or methodological differences between gene expression profiling platforms. This compendium approach allowed us to identify a specific and unique molecular transcript signature for this tumor compared to unrelated tumors. In essence, this approach enriches for changes specific to the patient's tumor and should thus represent relevant drug targets for therapeutic intervention. There were 3,064 differentially expressed genes (1,078 up-regulated, 1,986 down-regulated) in the lung tumor versus the blood and compendium. This analysis provided insight into those genes whose expression rate was likely to be a driving factor specific to this tumor, not identifying genes that correlate simply with proliferation and cell division. It is conceivable that such an approach, coupled with a greater understanding from multiple tumor datasets, could be replaced by the absolute quantification of oncogene expression as a means to determine clinical relevance.

5.2.1.5 Disease mechanism

I correlated mutated, amplified or differentially expressed genes with known cancer pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [310] and to drug targets present in the DrugBank database [144]. The 15 amplified, over-expressed or mutated genes in cancer pathways targetable by approved drugs are listed in Table 5.3. Some amplified genes, such as NKX3-1, RBBP8 and CABL1, were implicated in cancer but did not have well characterized functions or known drugs targeting them. I was not surprised to see LAMC1, a protein thought to be involved in metastasis, as the lung tumor had already metastasized from the tongue primary tumor.

The Ret proto-oncogene (RET) emerged as a particularly interesting gene, as it was both amplified and highly expressed. RET is a receptor tyrosine kinase that stimulates signals for cell growth and differentiation via the mitogen-activated protein kinase (MAPK)-extracellular signal-regulated kinase (ERK) pathway [311] and its constitutive activation is responsible for oncogenic transformation in medullary and papillary thyroid carcinoma [312]. In the lung tumor, RET was both highly amplified (hidden Markov model (HMM) level 4) and the most highly expressed known oncogene (34.5 fold change (FC) in lung relative to compendium; 123.2 FC in lung relative to blood) (Figure 5.3a). In addition, RET activating factors and MAPK pathway constituents were also highly expressed in the tumor. I also noted overexpression of the water channel protein Aquaporin-5 (AQP5), which has been implicated in multiple cancers and has been shown to activate Ras and its signaling pathways [313].

Aberrations leading to increased activation of the PI3K/AKT pathway are common in human cancers and are reviewed in [314]. Inactivating mutations and decreased expression (either by LOH or methylation) of PTEN, a tumor suppressor that reverses the action of PI3K, are the most frequently observed aberrations. Loss of PTEN expression has also been previously implicated in tongue squamous cell carcinoma [315]. In the patient tumor, PTEN was under-expressed (-109.7 FC in lung relative to compendium; -440.1 FC in lung relative to blood) and mapped to a region of heterozygous loss. Since PTEN mediates crosstalk between PI3K and RET signaling by negatively regulating SHC and ERK [316] and increased RET can also activate the PI3K/AKT pathway [311] (Figure 5.3a), loss of PTEN would up-regulate both the PI3K/AKT and RET-MAPK pathways, leading to decreased apoptosis, increased protein synthesis and cellular proliferation. In actuality, there was a LOH deletion in AKT1, under-expression of AKT2, mTOR, elF4E, and over-expression of the negative regulators eIF4EBP1 and NKX3-1. These changes appear to mitigate the effect of PTEN loss on the PI3K/AKT pathway and suggest that PTEN loss serves primarily to activate the RET pathway to drive tumor growth.

Like EGFR, RET also activates the RAS/ERK pathway (Figure 5.3a). Therefore, increased expression of RET provides a plausible explanation of the failure of the EGFR inhibitor erlotinib to control proliferation of this tumor. PTEN loss has also been implicated in resistance to the EGFR inhibitors gefitinib [317] and erlotinib [318]. Lastly, the mutated RB1 may also play a role in the observed erlotinib insensitivity, as the loss of both RB1 and PTEN as seen in this tumor has previously been implicated in gefitinib resistance [319].

As RET and PTEN were the most significant cancer-associated aberrations in the patient tumor, we performed FISH and immunohistochemical analysis on RET, PTEN, and RBBP8 to confirm their amplification statuses (Figure 5.4).

5.2.2 Therapeutic intervention

The integration of copy number, expression and mutational data generated a compelling hypothesis of the mechanism driving the tumor. This allowed us to identify drugs that target the observed up-regulation of the MAPK pathways through RET over-expression and PTEN deletion (Table 5.4). The approved cancer drugs sunitinib and sorafenib have wide kinase polypharmacologies (Appendix B), but were top candidates due their inhibition of RET. I then validated that other major protein targets of sunitinib and sorafenib, including RAFs, CSF-1R, FLT3, and VEGFRs, and PDGFRs, were expressed in the tumor and not mutated. We chose to administer sunitinib for three reasons: first, due to its wider range of kinase targets it may be able to concurrently target other pathways in the tumor cell; second, one of those targets (PDGFR) is highly expressed in the tumor and its activating factor PDGFB appears to be amplified; third, sorafenib also targets the MAPK pathway protein RAF and thus may be a viable treatment option in case the tumor develops resistance to sunitinib through a non-RAF-mediated change.

The patient gave his full and informed consent to initiate therapy with this medication and was fully aware that adenocarcinoma of the tongue was not an approved indication for sunitinib. Clinical administration of the RET inhibitor sunitinib showed evident shrinking of the patient tumors (Figure 5.5c), consistent with the hypothesis that RET-targeting drugs should inhibit the up-regulated MAPK proliferation pathway driving the tumor. The drug

was administered using standard dosing at 50 mg, orally, every day for 4 weeks followed by a planned 2 weeks off of the drug. After one such round of therapy, the patient had a PET-CT scan which was compared to the baseline pretreatment scan (Figure 5.5b). Using Response Evaluation Criteria in Solid Tumors (RECIST) criteria [320], the lung metastases had decreased in size by 22% and no new lesions had appeared. In contrast, the tumor exhibited 16% growth in the month pre-treatment (Figure 5.5b). However, due to typical side effects, the patient's sunitinib dose was reduced to 37.5 mg daily. Repeated scanning continued to show disease stabilization and the absence of new tumor nodules for four months.

5.2.3 Cancer recurrence

After 4 months on sunitinib, the patient's CT scan showed evidence of growth in the lung metastases. He was then switched to sorafenib and sulindac, another medication thought to be of potential benefit given his initial genomic profiling (Table 5.4). Within 4 weeks a CT scan showed disease stabilization and he continued on these agents for a total of 3 months when he began to develop symptoms of disease progression. At this point he was noted to have developed recurrent disease at his primary site on the tongue, a rapidly growing skin nodule in the neck, and progressive and new lung metastases.

5.2.3.1 DNA sequencing and mutation detection

A tumor sample was removed from the metastatic skin nodule and was subjected to both WTSS and genomic sequencing (Table 5.1). The four somatic changes identified in the pre-treatment tumor were detected, suggesting that the skin tumor was likely to have metastasized from the lung tumor. Nine new non-synonymous protein coding changes were detected that were not present within either the pre-treatment tumor or the normal DNA (Table 5.2). Reexamination of the sequence reads from the initial tumor analysis did not reveal the presence of any of these nine new mutated alleles even at the single read level.

5.2.3.2 Copy number analysis

Extensive copy number variations were observed in the post-treatment sample that were not present before treatment (Figure 5.2). In the tumor recurrence, 0.13% of the genome displayed high levels of amplification, compared to 0.05% in the initial tumor sample. Also,

24.8% of the initial tumor showed a copy number loss whereas 28.8% of the tumor recurrence showed such a loss. We identified eight regions where the copy number status changed from a loss to a gain in the tumor recurrence and twelve regions where the copy number changed from a gain to a loss. In addition, copy number neutral regions of LOH arose on chromosomes 4, 7 and 11.

5.2.3.3 Transcriptome analysis

There were 459 differentially expressed genes (385 up-regulated, 74 down-regulated) in the metastatic skin nodule versus the blood-and-compendium. Of these, 209 overlapped with the differentially expressed genes in the lung tumor versus blood-and-compendium set. In the skin metastasis relative to lung there were 6,440 differentially expressed genes (4,676 up-regulated, 1,764 down-regulated), reflective of the tremendous change the tumor underwent. Changes in expression in both the lung and skin metastases were significantly associated with copy number changes.

Overall, I found 23 amplified, over-expressed or mutated genes in cancer pathways targetable by approved drugs are listed in Table 5.5. It is interesting to note the presence of additional laminin and matrix metalloproteinase proteins, as they are known to have important roles in cancer cell metastasis [321, 322].

5.2.3.4 Disease mechanism

The cancer recurrence exhibited strong up-regulation of transcripts from genes in both the MAPK/ERK and PI3K/AKT pathways (Figure 5.4b). There are striking increases in expression of the receptor tyrosine kinases RET, EGFR, PDGFRB and their growth factor ligands GFRA1 (GDNF family receptor alpha 1), NRTN (neurturin), and EGF. Other genes within these pathways, such as AKT1, MEK1 and PDGFA, also appear amplified in copy number in the skin tumor compared to the lung tumor.

Taken together, these data suggest that the mechanisms of resistance to the RET targeting selective kinase inhibitors sunitinib and sorafenib are the up-regulation of the targeted MAPK/ERK pathway and the parallel PI3K/AKT pathway. It can be speculated that perhaps

only a cocktail of targeted drugs (such as to RET, EGFR, mTOR, AKT) would be able to affect the proliferation of the tumor cells. The further activation of RET and its downstream pathway in the post-treatment tumor suggests that the RET inhibitors are still necessary to treat the patient. Another drug candidate is the novel AKT inhibitor perifosine, which is has shown benefit in phase II clinical trials for multiple myeloma, chronic myeloid leukemia, and other hematological cancers [323]. The mTOR inhibitors rapamycin and everolimus have recently been approved for treatment of renal cell carcinoma after sorafenib and sunitinib failed to affect the disease [324]. However, it is important to note that these drugs only target the mTOR-RAPTOR protein complex and not the mTOR-RICTOR protein complex, and that the mTOR-RICTOR complex was prevalent in the skin tumor (mTOR FC 3.4 in skin tumor versus lung tumor, RICTOR FC 158.4 in skin versus lung tumor, and RICTOR versus RAPTOR FC 105 in the skin tumor). In addition, since RAF is now upregulated, it may be necessary to target a lower position in the MAPK pathway. At the moment, MEK inhibitors are either under development or in clinical trials - such as U1026 [325] and AZD6244 [326]. FDA approval of drugs like perifosine and AZD6244 would provide more rationally targeted therapeutic options for the patient.

Sunitinib resistance has been observed to be mediated by IL8 in renal cell carcinoma [327]. This is also reflected in the tumor data, where IL8 became highly over-expressed in the cancer recurrence (FC 861.1 in skin tumor relative to lung tumor). Though the mechanism of resistance is still unclear, IL8 has been observed to transactivate EGFR and downstream ERK, stimulating cell proliferation in cancer cells [328].

5.3 Discussion

High-throughput sequencing provided a comprehensive determination of copy number alterations, gene expression changes, and protein coding mutations in the patient's tumor. Correlation of the up-regulated and amplified gene products with known cancer-related pathways provided a putative mechanism of oncogenesis that was validated through the successful administration of targeted therapeutic compounds. Sequence analysis of the protein coding regions was also able to determine that the drug binding sites for known targets of sunitinib and sorafenib were intact.

Both sunitinib and sorafenib appeared to be viable options for the patient, and are widely used in sequence to treat renal cell carcinoma [329]. We chose to first apply sunitinib due to its wider polypharmacology (Appendix B), the apparent activation of PDGFR in the tumor, and the potential ability for sorafenib to treat sunitinib-resistant disease through RAF inhibition. However, several recent studies in metastatic renal cell carcinoma have hypothesized otherwise [330, 331]. They suggested that since sorafenib is a less potent inhibitor of certain kinases than sunitinib (Appendix B), the latter drug would be able to overcome resistance accrued during sorafenib treatment. This hypothesis is still under contention, as another large study in Czech patients did not observe any benefit for sorafenib-sunitinib therapy compared to sunitinib-sorafenib therapy [332]. In our case, sorafenib appeared to be effective for the patient after sunitinib resistance. The activation of PDGFRs may have contributed to the efficacy of sunitinib, and the presence of non-mutated RAF proteins in the skin metastasis suggest that sorafenib's potent RAF inhibition may have contributed to its therapeutic benefit (i.e. sunitinib treatment did not cause RAF mutations). In short, sequencing was also useful in helping us select the order of treatments.

The patient's initial tumor had molecular features previously implicated in other cancers. For instance, loss of copy number on 18q has been frequently observed in colorectal metastases. It is believed that these metastases are driven by inactivation of the tumor suppressor protein SMAD4 (on 18q) and the allelic loss of 18q [333]. The expression level of SMAD4 in the patient's tumor was found to be very low (43-fold lower than in samples within our compendium of tumor expression data); hence, down-regulation of SMAD4 along with loss of 18q also appear to be properties of the tumor. Another example is the amplification of RET, whose oncogenic transformation is known to drive medullary thyroid cancer, and play an important role renal cell carcinoma. These aberrations have not been previously observed in tongue cancers, and add insight into potential disease mechanisms for tongue adenocarcinomas.

The observation that the RET pathway had increased activity in the metastasis is important. First, it shows that resistance mechanisms may not always work by circumvention through

parallel pathways or acquired mutations, but can also further upregulate the targeted pathway. This leads to the second point, that in such cases the same drug (sunitinib/sorafenib) should not be discontinued in order to try a different chemotherapy. Typically, after one drug fails to stabilize the disease, it is stopped and another drug is tried as second-line therapy [332, 334]. However, our study suggests that sometimes the drug should be continued and at higher dosages if possible. Such a test can be incorporated into existing clinical and pathological tests to inform therapeutic options by testing if a known target of a working drug (i.e. RET in our case) has increased in expression. An interesting artifact of our analysis was the annotation of the anti-androgen bicalutamide as an inhibitor of IL6. Though bicalutimide was annotated in DrugBank 2.5 as an inhibitor of Interleukin-6, it has since been corrected and removed from the newer DrugBank 3.0. It is thus important to verify all important leads in the literature since database curation, like any other process, is prone to human error.

Though I have summarized the aberrations known to be involved in cancer pathways, many other changes accrued in the tumor during the 8 months of treatment. Just the expression analysis identified over 6000 differentially expressed genes in the post- versus pre-treatment tumor. Some of these may have contributed to the pathogenesis of the disease but are not currently known as being important in cancer. As our understanding of cancer biology is far from complete, it is possible that these drugs may have elicited the observed clinical benefit for reasons unrelated to the hypothesis. However, the skin metastasis showed changes corresponding to acquired resistance to RET-inhibition and suggests that the hypothesized disease mechanisms contributed, at least in part, to the tumor growth.

Our study provided clinically useful information and the rationale for a therapeutic regime that, whilst not curative, did establish stable disease for several months. We propose that complete genetic characterization in this manner represents a tractable methodology for the study of rare cancer types and can aid in the determination of relevant therapeutic approaches in the absence of established interventions. Furthermore, the establishment of repositories containing the genomic and transcriptomic information of individual cancers coupled with their clinical responses to therapeutic intervention will be a key factor in furthering the utility

of this approach. We eenvisage that as sequencing costs continue to decline, whole genome characterization will become a routine part of cancer pathology.

5.4 Methods

5.4.1 Sample preparation

Tumor DNA was extracted from formalin-fixed, paraffin-embedded lymph node sections (slides) using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Mississauga, ON, Canada). Normal DNA was prepared from leukocytes using the Gentra PureGene blood kit as per the manufacturer's instructions (Qiagen). Genome DNA library construction and sequencing were carried out using the Genome Analyzer II (Illumina, Hayward, CA, USA) as per the manufacturer's instructions. Tumor RNA was derived from fine needle aspirates of lung metastases and normal RNA was extracted from leukocytes using Trizol (Invitrogen, Burlington, ON Canada}) and the processing for transcriptome analysis was conducted as previously described [299, 335, 336]. The relapse sample was obtained by surgical excision of the skin metastasis under local anesthetic 5 days after cessation with sorafenib/sulindac treatment. DNA was extracted using the Gentra PureGene Tissue kit and RNA was extracted using the Invitrogen Trizol kit, and the genomic library and transcriptome library were constructed as previously described.

5.4.2 Mutational detection and copy number analysis

DNA sequences were aligned to the human reference, HG18, using MAQ version 0.7.1 [337]. To identify mutations and quantify transcript levels, WTSS data were aligned to the genome and a database of exon junctions [299]. SNPs from the tumor tissue whole genome shotgun sequencing and WTSS were detected using MAQ SNP filter parameters of consensus quality = 30 and depth = 8 and minimum mapping quality = 60. All other parameters were left as the default settings. Additional filters to reduce false positive variant calls included: the base quality score (MAQ qcal) of a variant had to be ≥20; and at least one-third of the reads at a variant position were required to possess the variant base pair. SNPs present in dbSNP [338] and established individual genomes [296, 339, 340] were subtracted as well as those detected in the normal patient DNA. SNPs present in the germline sample (blood) were detected using MAQ parameters at lower threshold of consensus quality = 10

and depth = 1 and minimum mapping quality = 20 in order to reduce false positive somatic mutations. Initially, non-synonymous coding SNPs were identified using Ensembl versions 49 and 50; the updated analysis presented here used version 52_36n. Candidate protein coding mutations were validated by PCR using primers using either direct Sanger sequencing or sequencing in pools on an Illumina GAiix. In the latter case, amplicons were designed such that the putative variant was located within the read length performed (75bp). For copy number analysis, sequence quality filtering was used to remove all reads of low sequence quality ($Q \leq 10$). Due to the varying amounts of sequence reads from each sample, aligned reference reads were first used to define genomic bins of equal reference coverage to which depths of alignments of sequence from each of the tumor samples were compared. This resulted in a measurement of the relative number of aligned reads from the tumors and reference in bins of variable length along the genome, where bin width is inversely proportional to the number of mapped reference reads. A HMM was used to classify and segment continuous regions of copy number loss, neutrality, or gain using methodology outlined previously [341]. The sequencing depth of the normal genome provided bins that covered over 2.9 gigabases of the HG18 reference. The five states reported by the HMM were: loss (1), neutral (2), gain (3), amplification (4), and high-level amplification (5). LOH information was generated for each sample from the lists of genomic SNPs that were identified through the MAQ pipeline. This analysis allows for classification of each SNP as either heterozygous or homozygous based on the reported SNP probabilities. For each sample, genomic bins of consistent SNP coverage are used by an HMM to identify genomic regions of consistent rates of heterozygosity. The HMM partitioned each tumor genome into three states: normal heterozygosity, increased homozygosity (low), and total homozygosity (high). We infer that a region of low homozygosity represents a state where only a portion of the cellular population had lost a copy of a chromosomal region.

5.4.3 Gene expression analysis

Transcript expression was assessed at the gene level based on the total number of bases aligning to Ensembl (v52) [342] gene annotations. The tumor transcriptome library was found to be enriched for fragments representing contaminating genomic DNA, which was compensated for by performing a genomic subtraction. We estimated 84% genomic

contamination based on the proportion of intergenic reads present in the tumor library. We then compensated for the contamination by subtracting, for each gene, the expected coverage from genomic contamination.

The corrected and normalized values for tumor gene expression (both skin and lung metastases) were then used to identify genes differentially expressed with respect to the patient's germline (blood) and a compendium of 50 previously sequenced WTSS libraries. This compendium was composed of 19 cell lines and 31 primary samples representing at least 19 different tissues and 25 tumor types as well as 6 normal or benign samples (Appendix C). Tumor versus compendium comparisons used outlier statistics and tumor versus blood used Fisher's exact test. We first filtered out genes with less than 20% non-zero data across the compendium. This was necessary to avoid cases where a small expression value in the tumor receives an inflated rank when all other libraries reported zero expression (a problem common to sequencing-based expression techniques when libraries have insufficient depth). Next, we defined over-expressed genes as those with outlier and Fisher P-values < 0.05 and FC for tumor versus compendium and tumor versus blood > 2 and > 1.5, respectively. Similar procedures were used to define under-expressed genes. In addition to lung/skin metastasis versus compendium/normal blood we also compared the skin and lung metastases directly. P-values for differential expression were corrected with the Benjamini and Hochberg method [343]. Overlaps were determined with the BioVenn web tool [344].

5.4.4 Immunohistochemistry

Immunohistochemistry was performed using automated methods as previously described [345], with the following antibodies: monoclonal rabbit anti-human PTEN 1:25 dilution (clone 138C6, cat# 9559, Cell Signaling Technology, Beverly, MA), goat polyclonal anti-human RET diluted 1:25 dilution (clone C-20, cat# sc-1290, Santa Cruz Biotechnology, Santa Cruz, CA), monoclonal rabbit anti-human NTRK1 1:350 dilution (clone 14G6, cat# 2508, Cell Signaling Technology), and undiluted CONFIRM anti-human EGFR (clone 3C6, cat# 790-2988, Ventana, Tucson, AZ). Hematoxylin and eosin staining is performed using standard reagents and methods.

5.4.5 Fluorescence in situ hybridization

Bacterial artificial chromosomes (BACs) were obtained from the Children's Hospital Oakland Research Institute (Oakland, CA). The BACs RP11-124O11, labelled with SpectrumRed (Abbott Molecular, Abbott Park, IL), and RP11-348I3, labelled with SpectrumGreen, flanked the RET locus and detect disruption of RET. BACs RP11-66D17 (red) and RP11-1038N13 (green) flanked the NTRK1 locus and detect disruption of NTRK1. BAC RP11-104H10 (red) was used to detect RBBP8 copy number. The PTEN and EGFR loci were detected with commercial probes (EGFR: Vysis LSI EGFR SpectrumOrange/CEP 7 SpectrumGreen probe, cat# 32-191053; PTEN: Vysis LSI PTEN Spectrum Orange/CEP 10 SpectrumGreen dual color probe, cat# 32-231010; Abbott Molecular). Commercial centromeric probes for chromosomes 10 and 18 were used in conjunction with the RET and RBBP8 BAC probes, respectively (chr. 10: CEP 10 SpectrumAqua, cat# 32-131010; chr. 18: CEP 18 (D18Z1) SpectrumAqua, cat# 32-131018). FISH was performed as previously described [346].

Figure 5.1 Timeline of treatment and summary of sampled sites for the patient

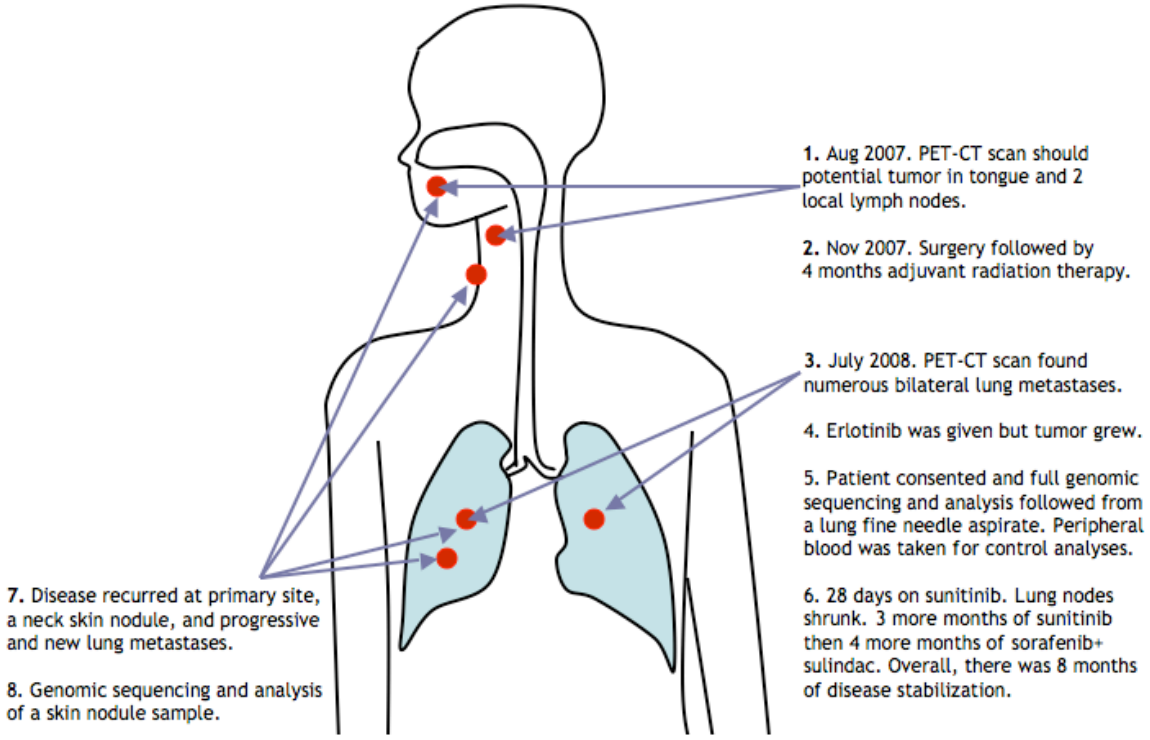


Figure 5.2 Identified regions of chromosomal copy number variation (CNV) and loss of heterozygosity (LOH).

Regions in the pre-treatment (T1 – lung tumor) and post-treatment (T2 – skin tumor) tumor samples and matched normal patient DNA (R - reference) plotted in Circos format [347]. CNV values are the hidden Markov model (HMM) state. LOH values are shown in the shaded green track. Δ indicates the degree in change of HMM or LOH state between the two cancers.

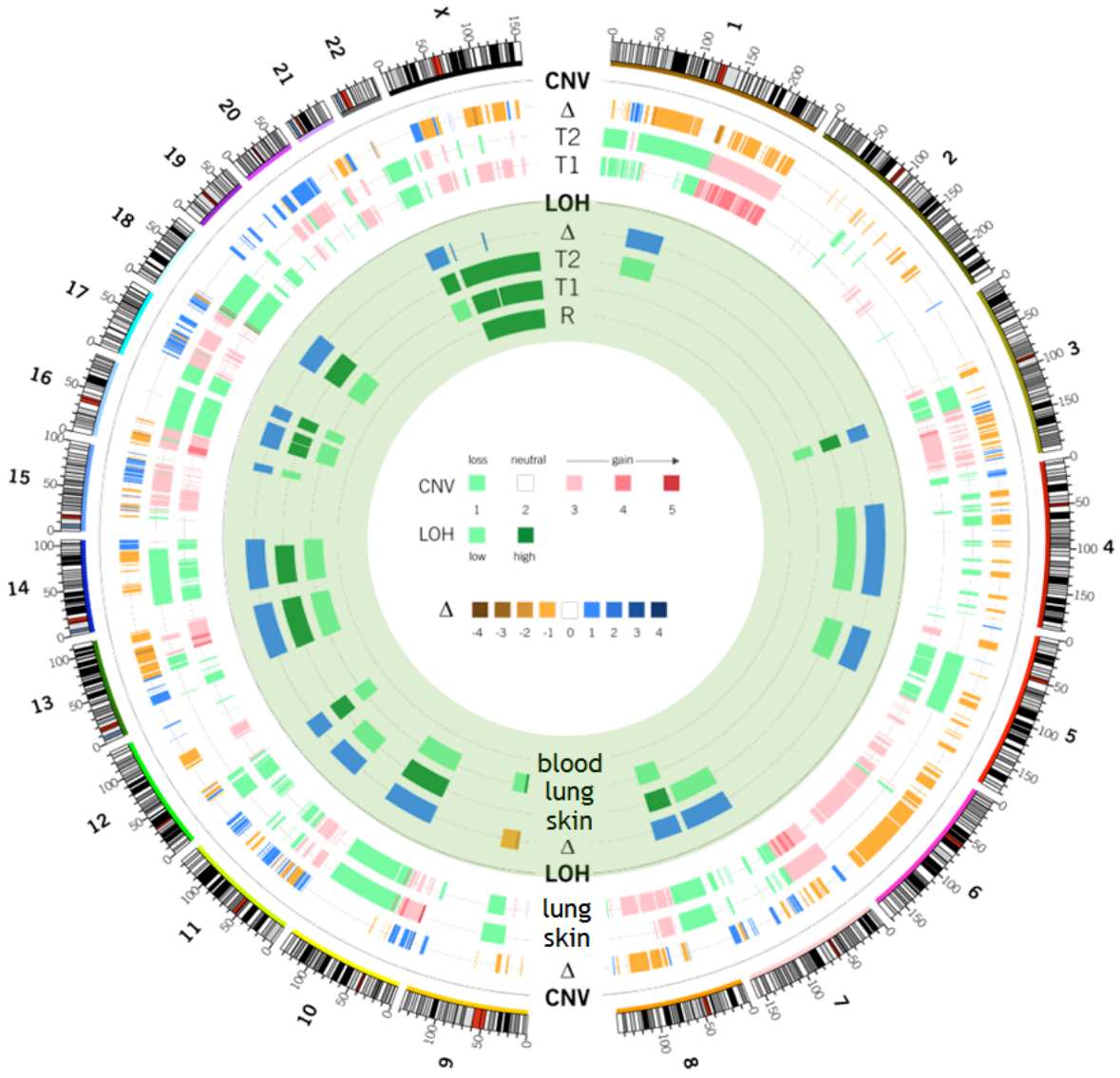
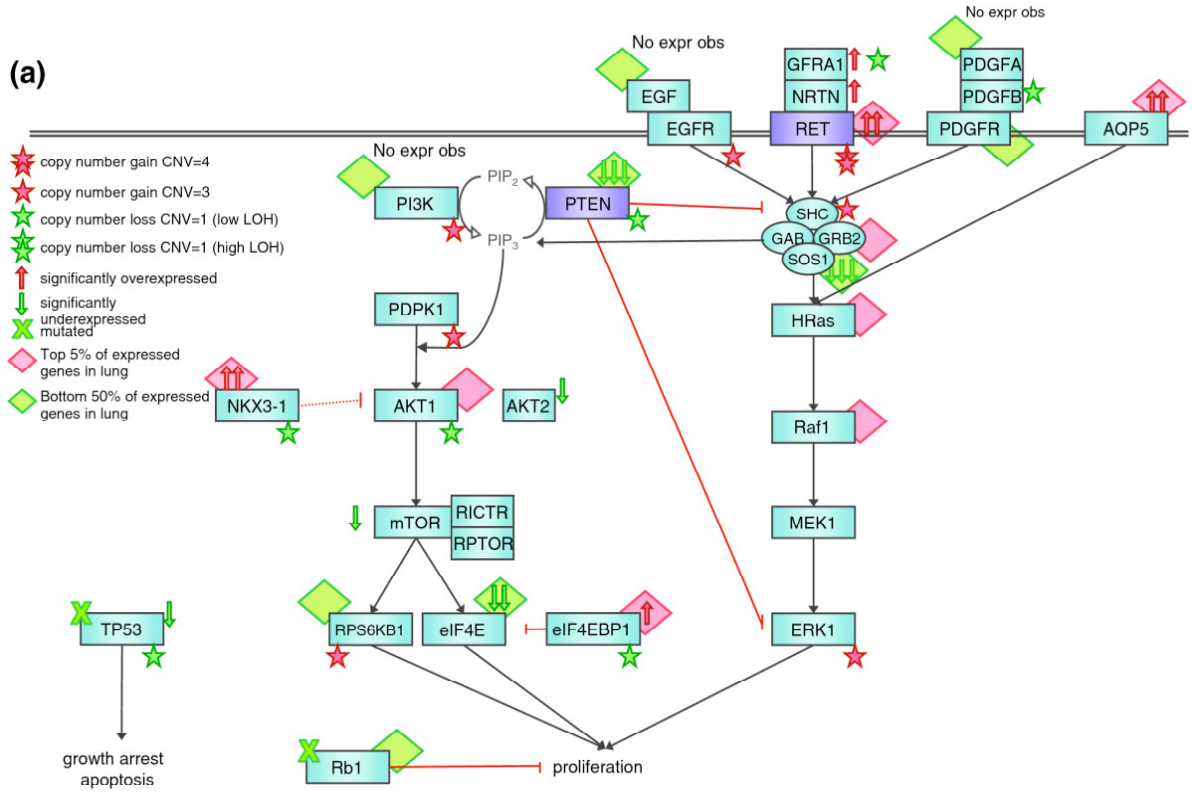


Figure 5.3 Cancer signaling pathways affected within the tumor.

(a) Pre-treatment: overall, the down-regulation of PTEN and up-regulation of the RET signaling pathway appear to be driving tumor proliferation. Increased signaling independent of EGFR is consistent with the observed erlotinib insensitivity of the tumor. The number of arrows denoting significantly over- or under-expressed genes are quantified using fold change of tumor versus compendium in (a), and primary tumor versus the tumor recurrence in (b): 1 arrow is FC ≥2; 2 arrows is FC ≥10; and 3 arrows is FC ≥50. CNV, copy number variation.



(b) Post-versus pre-treatment: after treatment with the RET inhibitors sunitinib and sorafenib. There is a marked increase in the signaling of pathway constituents leading to tumor proliferation. Black and red pathway arrows represent activation and inhibition, respectively. Dotted arrows represent indirect interactions.

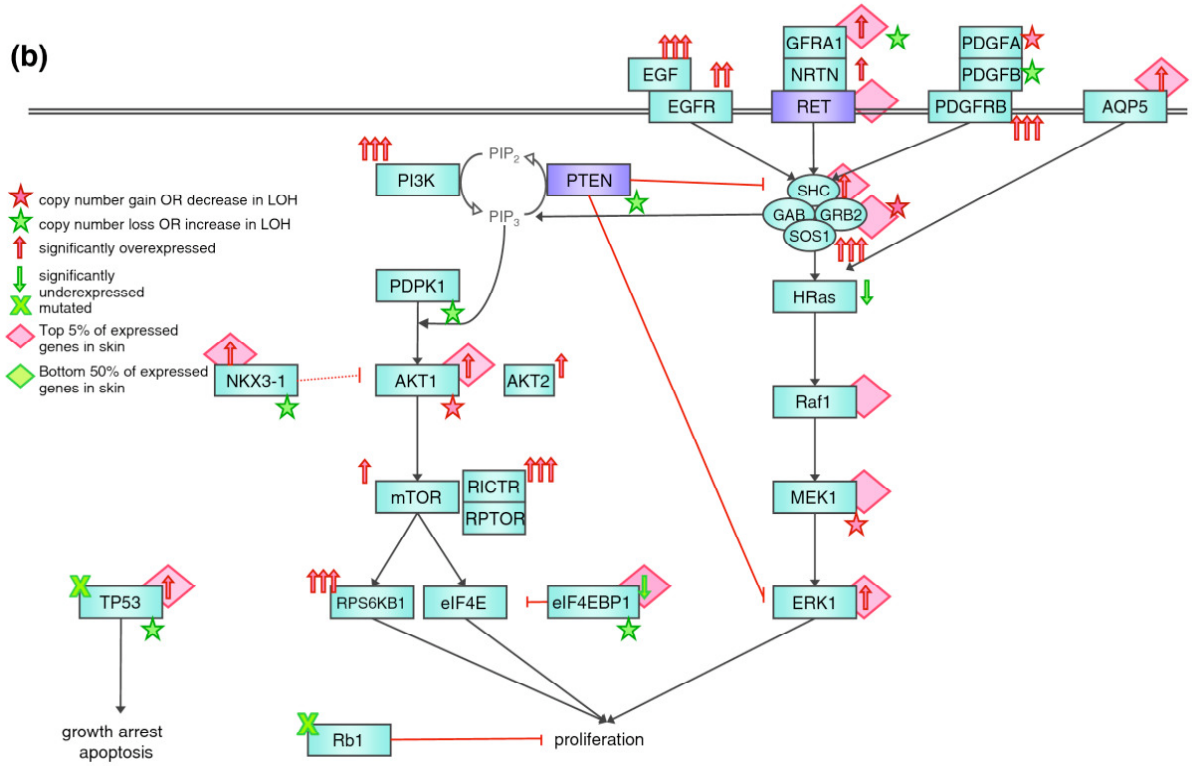


Figure 5.4 Fluorescent in situ hybridization (FISH) and immunohistochemical analysis of the sublingual adenocarcinoma.

(a) Hematoxylin and eosin stained section of tumor (20× objective). (b) Striking amplification of RBBP8 (40×, with RBBP8 probe in red). (c) Focal nuclear and cytoplasmic expression of PTEN (20×) is associated with (d) a missing red signal indicating monoallelic loss of PTEN (100×; the orange gene-specific probe signals are decreased in number compared to the centromeric probe). (e) Diffuse, strong cytoplasmic expression of RET (20×) is associated with (f) amplification of the RET gene (40× with bacterial artificial chromosomes flanking the RET gene labeled in red and green).

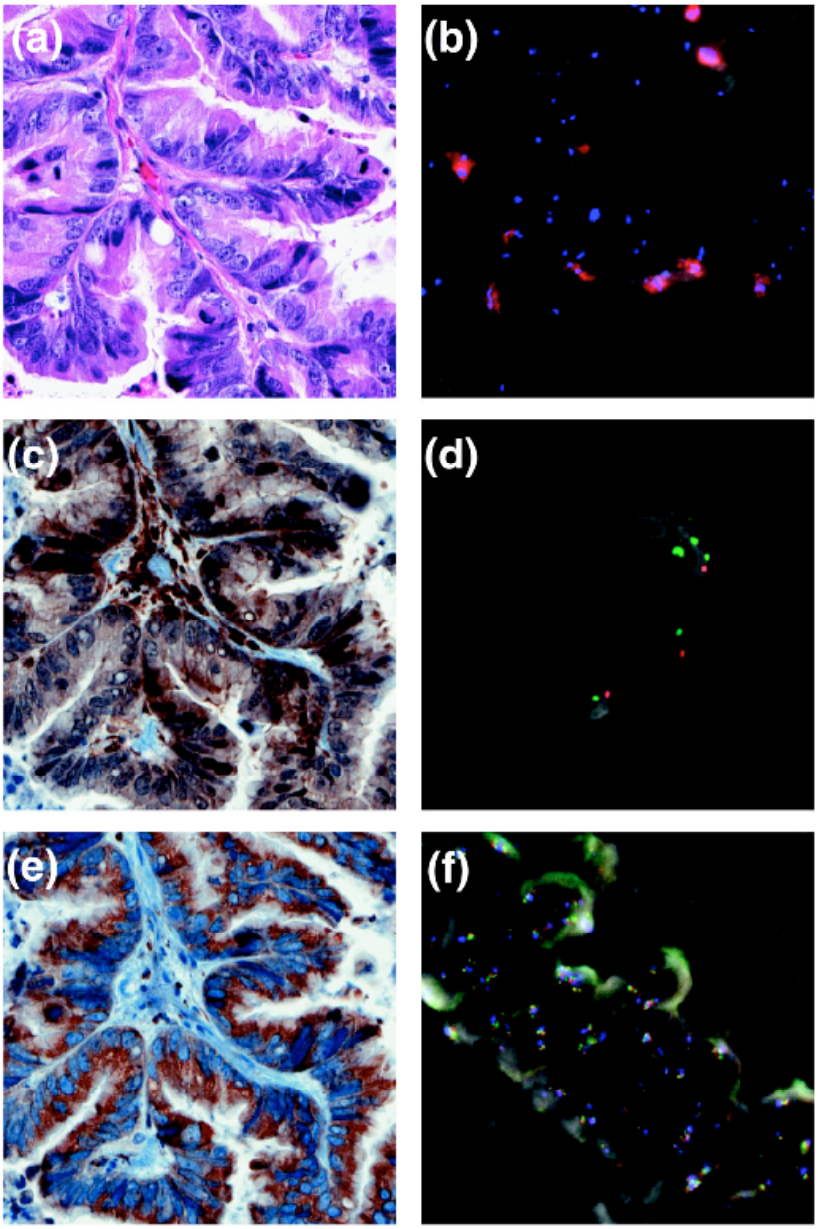


Figure 5.5 PET-CT scans of the patient.

(a) 1 October 2008, 1 month before sunitinib initiation. (b) 29 October 2008, baseline before sunitinib initiation on 30 October 2008. (c) 9 December 2008, 4 weeks on sunitinib.

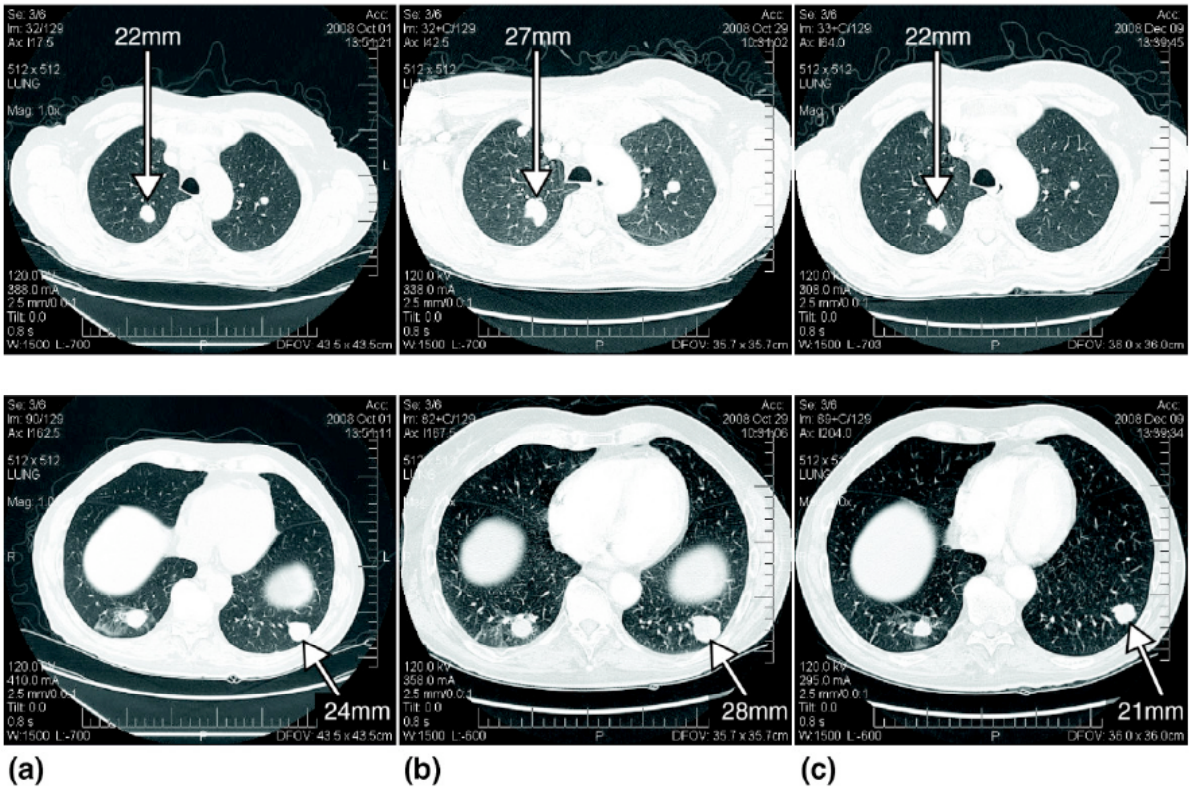


Table 5.1 Summary of tumor and normal samples sequenced in the study.

Lung Tumor	DNA	lymph node biopsy sections (slides)	2,584,553,684 42-bp reads
Lung Tumor	RNA	fine needle aspirates of lung biopsy	498,229,009 42-bp reads
Lung Normal	DNA	leukocytes (peripheral blood)	342,019,291 42-bp reads
Lung Normal	RNA	leukocytes (peripheral blood)	62,517,972 42-bp reads
Skin Tumor	DNA	skin nodule	1,262,856,802 50-bp reads
Skin Tumor	RNA	skin nodule	5,022,407,108 50-bp reads

Table 5.2 Predicted protein coding somatic changes within the initial tumor (T1) and the drug resistant recurrent tumor (T2)

Validated non-synonymous single nucleotide variations (SNVs) predicted by high-throughput sequencing are listed with the corresponding chromosome position (Chr. position), Ensembl gene ID, the base at this location in the reference genome (Ref.), the observed base (Obs.), the amino acid change as a result of the SNV, and the Ensembl description for this gene. The first four SNVs marked T1 were identified in the primary tumor and were validated using PCR and Sanger sequencing on germline and tumor genomic DNA. The remaining nine SNVs marked T2 were identified in the post-treatment secondary tumor and were validated by Illumina sequencing. SNVs in the initial tumor were also identified and validated in the recurrent tumor.

Tumor	Chr. position	Ensembl gene ID	Ref.	Obs.	Protein change	Description
T1	6: 28352058	ENSG00000197062	G	T	G62C	Zinc finger. SCAN domain - containing protein 26
T1	8: 106884238	ENSG00000169946	A	G	K785E	Zinc finger protein multitype 2. Friend of GATA protein 2 (FOG-2)
T1	13: 47832247	ENSG00000139687	T	A	L234*	Retinoblastoma-associated protein (pRb)
T1	17: 7518231	ENSG00000141510	C	A	D259Y	Tumor suppressor *P53*
T2	1: 35608585	ENSG00000146463	G	C	Q317H	Zinc finger protein 262
T2	2: 196431742	ENSG00000118997	C	G	V2590L	Dynein heavy chain 7, axonemal Ciliary dynein heavy chain 7) (HDHC2)
T2	4: 78747983	ENSG00000156234	G	A	R56H	B cell-attracting chemokine 1 (BCA-1)
T2	6: 33281235	ENSG00000204228	G	A	A141T	Estradiol 17-beta-dehydrogenase 8 (17-beta-HSD 8) (Ke-6)
T2	7: 82419723	ENSG00000186472	T	C	T2759A	Protein piccolo (Aczonin)
T2	11: 105355581	ENSG00000152578	C	T	R872C	Glutamate receptor 4 Precursor (GluR4) (Glutamate receptor ionotropic, AMPA 4)
T2	14: 19414855	ENSG00000165762	C	T	L197F	Olfactory receptor 4K2
T2	14: 63500386	ENSG00000054654	C	G	A302G	Nesprin-2 (Nuclear envelope spectrin repeat protein 2) (Syne-2)
T2	18: 8333477	ENSG00000173482	G	A	A929T	Receptor-type tyrosine-protein phosphatase mu Precursor (R-PTP-mu)

Table 5.3 Cancer related observed lung tumor aberrations.

Proteins that are amplified compared to blood, significantly overexpressed compared to both blood and compendium, or mutated. Highly amplified refers to an HMM classification value of 4 and amplified to an HMM value of 3. Only proteins that are known to be targets of approved drugs are listed. The last column lists a few approved drugs that are annotated in DrugBank as binding to each target.

Target	**Target name**	**Genome aberration in lung tumor**	**Approved drug**
RET	Proto-oncogene tyrosine-protein kinase receptor ret	significantly over expressed highly amplified	sunitinib sorafenib
EGLN1	Egl nine homolog 1	highly amplified	vitamin C
LAMC1	Laminin subunit gamma-1	highly amplified	alteplase reteplase
PTGS2	Prostaglandin G/H synthase 2	highly amplified	etoricoxib carprofen
BMP2	Bone morphogenetic protein 2	amplified	simvastatin
CYCS	Cytochrome c	amplified	minocycline melatonin
EGFR	Epidermal growth factor receptor	amplified	gefitinib erlotinib
GSK3B	Glycogen synthase kinase-3 beta	amplified	lithium
HDAC2	Histone deacetylase 2	amplified	vorinostat
IL6	Interleukin-6	amplified	bicalutamide arsenite
MAPK3	Mitogen-activated protein kinase 3	amplified	sulindac isoprotenerol
NTRK1	High affinity nerve growth factor receptor	amplified	imatinib
PRKCB	Protein kinase C beta type	amplified	vitamin E
RAC1	Ras-related C3 botulinum toxin substrate 1	amplified	simvastatin
RXRG	Retinoic acid receptor RXR-gamma	amplified	tretinoin adapalene

Table 5.4 Potential therapeutics targeting the observed lung aberrations.

Drug	Known mechanism & indications	Targeted aberrations
Sunitinib	Targets PDGFRs, VEGFRs, RET, KIT, CSF1R, FLT3. Approved for GIST and RCC. In trials for thyroid cancer.	Up-regulation the MAPK pathway increases cell proliferation. RET, a validated thyroid cancer target, and its growth factors are amplified and overexpressed AQP5 a known activator of this pathway is overexpressed MAPK3 (ERK1) is amplified. BRAF is a target in thyroid cancer. PTEN, a suppressor of this pathway, is highly down-regulated.
Motesanib	Targets VEGFRs, PDGFRs, KIT, RET. In trials for thyroid cancer, GIST, NSCLC.	
Sorafenib	Targets BRAF, RAF1, RET, VEGFRs, PDGFRB, KIT, FLT3. Approved for RCC and HCC. In trials for thyroid cancer.	
Sulindac	An NSAID COX inhibitor for inflammation but also inhibits MAPK3 (ERK1).	

Table 5.5 Cancer related observed skin tumor aberrations.

Proteins that are amplified compared to blood, significantly overexpressed compared to both blood and compendium, or mutated. Highly amplified refers to an HMM classification value of 4 and amplified to an HMM value of 3. Only proteins that are known to be targets of approved drugs are listed. A few approved drugs known to inhibit each target are listed.

Target	Target name	Genome aberration in skin tumor	Approved drug
RET	Proto-oncogene tyrosine-protein kinase receptor ret	significantly over expressed highly amplified	sunitinib sorafenib
AKT1	RAC-alpha serine/threonine-protein kinase	significantly over expressed	arsenite
BMP2	Bone morphogenetic protein 2	amplified	simvastatin
CYCS	Cytochrome c	amplified	minocycline melatonin
EGFR	Epidermal growth factor receptor	amplified	gefitinib erlotinib
EGLN1	Egl nine homolog 1	amplified	vitamin C
ERBB2	Receptor tyrosine-protein kinase erbB-2	amplified	lapatinib
GRB2	Growth factor receptor-bound protein 2	amplified	pegademase bovine
GSK3B	Glycogen synthase kinase-3 beta	amplified	
IL6	Interleukin-6	amplified	bicalutamide arsenite
ITGA2B	Integrin alpha-IIb	amplified	tirofiban
LAMA1	Laminin subunit alpha-1	amplified	alteplase reteplase
LAMC1	Laminin subunit gamma-1	amplified	alteplase reteplase
MAPK3	Mitogen-activated protein kinase 3	amplified	sulindac isoprotenerol
MMP9	Matrix metalloproteinase-9	amplified	minocycline simvastatin
NTRK1	High affinity nerve growth factor receptor	amplified	imatinib
PRKCA	Protein kinase C alpha type	amplified	vitamin E
PRKCB	Protein kinase C beta type	amplified	vitamin E
PTGS2	Prostaglandin G/H synthase 2	amplified	etoricoxib carprofen
RAC1	Ras-related C3 botulinum toxin substrate 1	amplified	simvastatin
RARA	Retinoic acid receptor alpha	amplified	isotretinoin alitretinoin
RXRG	Retinoic acid receptor RXR-gamma	amplified	tretinoin adapalene
STAT5B	Signal transducer and activator of transcription 5B	amplified	dasatinib

6 Conclusions and Future Directions

Drug repositioning has become increasingly studied in recent years due to the consistently low rate of new drug approvals. Given the vast number of potential drug-target interactions, computational methods are a valuable parallel approach to experimental methods. The primary goals of this thesis were to find novel drug repositioning candidates by (1) developing a computational method to predict novel drug-target interactions and (2) better understanding of disease mechanisms. In Chapter 1, I reviewed the existing computational drug repositioning approaches.

6.1 Molecular docking to find novel drug-target interactions

In Chapter 2, I described my computational repositioning approach, involving the largest molecular cross-docking study to date. Molecular docking is a more realistic model of binding interaction compared to existing methods based on protein sequence, protein structure, or chemical similarity, and can detect drug repositioning candidates that are not structurally similar to existing drugs. Conceptually, my cross-docking approach also models a more realistic biological environment - once a drug enters a cell, which of the multitude of different proteins will it bind to? In the future, it could be informative to classify target proteins by their subcellular localization in order to better model the cellular environment.

A major limiting factor of large-scale docking studies is the requisite computational power to virtually screen millions of interactions. Therefore, the results and analysis of this study, taking over 3 weeks on a 1000-processor cluster, will be an important contribution to the molecular docking community. However, without further processing, these results would consist of mainly false positive interactions. Using software-standard docking score thresholds, I would have predicted over 100,000 interactions, containing only 1.1% known interactions. Experimentally validating all the predictions would be infeasible.

The computational goal of my method was develop methods to filter out as many false positive interactions as possible. The ranking and scoring criteria I developed allowed for a more rational selection of top interactions and enriched the predicted set for known

interactions several hundred fold. My filtering method also differed from existing machine learning approaches in that there was no need to train on known binders for each target. The high EFs suggested that the predicted interactions were more likely to be true binders compared to predictions from other thresholds, and would be more efficient for experimental validation. Indeed, I found literature validation for 31 of the top predictions that were not annotated in DrugBank. Since it is infeasible for drug-target interaction databases to manually curate all the literature for all targets and drugs, I suggest that virtual screening studies can also aid in annotating existing drug-target interaction databases.

My method is scalable to larger datasets - as long as the docking can be feasibly completed. The set of proteins can be increased as new structures are entered in PDB. Moreover, homology models of proteins can be included provided that these structures pass the reliable-for-docking criteria. As docking and scoring mechanisms continue to improve, it is foreseeable that more known interactions will be docked well, also leading to an increase in the number of reliable-for-docking targets in the dataset.

Current docking methods are recognized to have many weaknesses: lack of protein flexibility, lack of solvent molecules, poor scoring functions, to name a few. I countered these shortcomings by using 'reliable' proteins for which at least one known drug could be docked well, hypothesizing that for these proteins the shortcomings did not override the predictive ability of the docking. Ideally, molecular dynamics would be the most realistic simulation of protein-ligand binding in 3-dimensional space; however, it is currently not feasible to perform large-scale molecular docking in an automated fashion or in a timely manner.

In this chapter, I focused on finding novel targets for existing drugs, in order to determine novel therapeutic indications, added insight into mechanism of action, and better understanding of adverse reactions at the molecular level for these drugs. However, this method is not limited to only approved drugs, and can be used to select top predictions from any virtual screening study; for example, I could screen 20 million PubChem compounds against a target, use the consensus score threshold to select the top 1000 compounds, then

dock these against the reliable target set to determine protein ranks and further filter top predictions. In short, molecular docking is a powerful method for determining protein-drug interactions and the docking approach and scoring thresholds I developed can be applied to improve any future docking prediction, whether for novel compounds or approved drugs.

6.2 In-depth docking of EGFR kinase to find novel repositioning candidates

In Chapter 3, I presented a detailed computational analysis to find inhibitors of the well-established drug target EGFR. The docking approach in Chapter 2 appeared to perform well for EGFR, greatly enriching the predicted set of inhibitors for known interactions (15 fold better than the standard software threshold). Compared to previous EGFR docking studies, my study was unique in that it used 23 crystal structures of the protein instead of just one. I hypothesized that this would allow us to better model protein flexibility. In addition, the protein rank criteria allowed us to further eliminate potential false positive predictions. Though a few true positive interactions were also filtered out, these were broad-spectrum binders of EGFR (ATP, ADP, staurosporine) that were not interesting drug candidates.

The anti-HIV drug TDF showed promise in cell line and Western blot assays, behaving similarly to the known EGFR drug gefitinib. However, TDF did not inhibit EGFR in direct binding assays, suggesting that the drug works through other mechanisms. I speculated that the drug may be acting through other ERBB family kinases that are similar to EGFR (but are not included in my database), or the other docking-predicted targets PIMT or GALE, or kinase proteins in between EGFR and ERK in the signaling pathway. More analysis will be needed to investigate these potential interactions using computational 3D structures (building homology models if possible) and experimental assays.

This study showed some of the limitations of my virtual screening approach. As mentioned for Chapter 2, not having true targets of TDF in the dataset will cause other false positive targets to pass protein rank thresholds. In addition, the scoring functions are still inaccurate and can cause true targets to score poorly and false targets to score well. These limitations are applicable to any drug-target prediction study using docking methods. However, the number of structures solved in PDB is rapidly increasing and as the scoring functions

improve, I believe that redoing this analysis in the future will provide a more informative set of predictions. In addition, as more EGFR crystal structures are deposited into PDB, they can also be added to the set of targets.

From this study, I have learned that experimental determination of drug-target binding also has many limitations. First, the assay conditions and experimental design can affect whether a direct binding interaction is detected and at what strength. Examples have been shown throughout the thesis, whether with previous nilotinib-MAPK14 results, gefitinib targets determined from two different studies, the BIM-8 PIM-1 interaction, the importance of ATP and FBS concentrations. In addition, observed inhibition in cell proliferation and signaling assays may have arisen through any number of targets, and may not be due to inhibition of the desired target. However, cell line assays may be the only option when direct binding assays are not readily available.

The inverse docking of TDF against the protein database aided in selecting top interaction predictions. However, it also allowed us to explore the potential polypharmacology of the drug through a drug-target network. The clinical utility of multi-targeting or 'dirty' drugs has been proven with cancer drugs like sunitinib and sorafenib. However, not knowing the contributions of each target means that there will still be more adverse effects than I would like. As more is understood about cancer pathways, an important milestone of drug discovery would be to design drugs that only inhibit specific combinations of targets, for greater efficacy and better safety profiles. The drug-target network I built from docking is one computational approach this problem, by allowing us to identify combinations of proteins that can be targeted by a single agent. Though targets were not similar in sequence or structure, like GALE and EGFR, some molecular properties in their binding sites allow them to dock well with the same drug. This network also suggests non-kinase drug targets that could be targeted in combination with EGFR.

6.3 Combining VS and HTS to find novel drug repositioning candidates for RSK

Based on my experience docking to EGFR, I made several amendments to my screening strategy. I first built homology models of RSK to supplement existing PDB structures, which either had structure gaps near ligand binding sites or were not in a peptide-bound conformation. Though it can be argued that docking to homology models is less reliable than docking to an original 3D structure, several studies have shown that models can be just as informative or even more informative when used in docking. Oshiro *et al*. studied the cyclin-dependent kinase CDK2 using templates with sequence identity of 43% to 60% and the DOCK software, and found that they could enrich the predicted binders by 3.5 or 3.1 times, respectively, using homology models [348]. This was only slightly lower than the 4.5 times enrichment obtained using actual CDK2 crystal structures, demonstrating the utility of homology models for VS. I found adding the two RSK homology models beneficial to my analysis, both in docking known inhibitors (Table 4.3) and docking HTS hits (for the 32 HTS hits, 5 docked best to a RSK model structure instead of an existing RSK crystal structure).

One of the reasons I constructed a peptide-bound model of RSK was to search for novel protein-protein (peptide pocket) small molecule inhibitors. To date, finding small molecule inhibitors of protein-protein interaction (PPI) interfaces remains a widely studied yet largely undeveloped area of drug discovery. Being able to target these interfaces would greatly increase the number of druggable targets for therapeutic purposes, since PPIs occur in all major biological and disease pathways. In particular, kinase targets share similar ATP-binding sites, so an additional peptide-binding site inhibitor could add a layer of specificity to kinase drugs. However, due to the different natures of small molecule and PPI sites, these types of inhibitors are extremely challenging to detect using existing VS and HTS methods [23]. In this study, the challenge presented as not having known RSK PPI inhibitors for positive control docking. Overall, Prestwick drugs had poor pmf- scores and none of the top predictions exhibited activity in the HTS screen. Conversely, none of the 32 HTS hits docked well to the peptide pocket. Only menadione may have bound near the peptide site, since it inhibited RSK-S6K but not RSK-YB-1 activity. Two likely factors emerged when comparing the HTS and VS results: first, the small Prestwick library may not have contained any strong RSK PPI inhibitors; and second, empirical scoring algorithms (like ICM) were unable to

score PPI inhibitors well since they were trained on existing PDB structure complexes (none of which involve PPI inhibitors). I thus believe that more PPI-specialized VS and HTS methods need to be developed before computationally determining PPI inhibitors becomes routinely feasible.

The second amendment to my strategy was to perform a much more in depth docking analysis of known inhibitors. The EGFR known inhibitors were very similar to ligands in PDB structure complexes; thus, the positive controls in that study were all cognate dockings. Here, I chose ten chemically diverse inhibitors of RSK, six of which represented the non-cognate docking scenario. Validating their scores, ranks, and known binding domains (NTKD/CTKD) allowed us to be more confident in my positive control analysis. This step also provided the predicted binding conformation of several RSK inhibitors that have not yet been solved in RSK crystal structures. These binding modes will be useful for understanding drug binding chemistries and designing derivative compounds.

PDB structures are often considered gold standard for docking purposes, as it is the central repository of protein three dimensional crystal structures. However, it is not often recognized that PDB structures often have errors of their own – whether human error or refinement error (i.e. incorrect sequence mapping errors [349]). Docking software like ICM attempt to adjust for many of these errors but cannot account for all cases. Other software should also be tried, as they would adjust for a different set of errors. Through the known inhibitor analysis, I was able to detect RSK structures not suitable for ICM docking. I thus found it important to examine the docking of each RSK structure individually, instead of grouping them together as in previous chapters.

The most significant change to my strategy was the inclusion of HTS results, parallel to the VS results. I was surprised to find that 32 of 1,120 off-patent drugs had RSK inhibitory activity. Of course, some drugs may have been false positive or weak inhibitors. However, if combinations of weak drugs are taken, the additive effect on RSK inhibition could become significant. Furthermore, these drugs would be chosen such that they would not share many other targets in common, aside from RSK. The overall effect would therefore be a strong

added RSK inhibition with a weak inhibition of other targets, resulting in milder adverse effects. While an attractive endeavour in theory, such combinations are challening to formulate in practice. Drug-drug interactions will need to be carefully understood, to ensure that drugs in the combination do not interfere with each others' absorption, distribution, metabolism, excretion, or actual clinical effect [350]. In addition, it would be important to ensure that weakly inhibited targets do not cooperate synergistically to cause adverse effects.

Aside from allowing us to select top Prestwick drugs for secondary screening, the combined screening approach also allowed a direct comparison of HTS and VS for the same chemical library. While the ability of docking to enrich for known RSK binders was very clearly established (PPV=23%, EF=26.1), I was interested to see how the scoring and ranking criteria enriched for HTS hits. There was very little overlap between VS and HTS results, with only 6 drugs in common between the top ~30 interactions from each analysis. However, many of the drugs passing the consensus score threshold included the HTS hits with the strongest activity (kaempferol, myricetin, ellipticine). Examining the drugs eliminated by visual criteria showed that rigid molecules like ellipticine formed few hydrogen bonds but could induce protein to conform during binding (like staurosporine). The new generation of flexible-protein docking algorithms may be able to improve detection of such compounds.

The EFs in this study were lower compared to the EGFR analysis, due to the smaller database size; however, the PPV rates remained around 20%. The EF analysis showed that VS allows us to screen only 19 compounds (<2% of the library) yet find six RSK inhibitors, allowing for a much more efficient experimental output in the absence of an HTS screen. In addition, VS methods were able to aid in identifying several HTS false positive hits, as was seen with menadione and the steroid hormone compounds.

Lastly, during the validation steps, we made sure to confirm a strong dose-dependent inhibition of RSK before continuing on to signaling or growth inhibition experiments. This prevented any TDF-like scenarios from occurring – because compounds that inhibit growth and signaling may be acting through any number of pathways. Even in the current results, the five follow-up drugs may be acting through a number of kinase proteins, such as EGFR or

ERK, which are above RSK in the MAPK signaling chain. A glimpse of other potential targets for each drug can be obtained from the protein rank analysis, when the top ~110 drugs were docked to the 252 drug-target database. As an example, myricetin with protein rank 75 had strong RSK inhibition but its poor protein rank indicated that it may also target many other proteins in the cell.

The behaviour of myricetin underscores one disadvantage for both HTS and VS screens: they are not reflective of the cellular environment. The dynamics of the target protein, the concentration at which the protein is present in the cell, local environments in the cell (including the subcellular localization) and presence of cofactors, chaperones, or competitors, may all affect the relevance of the HTS or VS result. Similarly, cellular assays are not necessarily reflective of results from an animal model study, and the drug efficacy and toxicity seen in an animal model may not translate to the clinic. In our computational method, we tried to account for cellular competitors by including a host of other drugs and target proteins in the docking process. We have identified drugs that show direct inhibition of RSK and cellular inhibition of TNBC cell lines. The next step thus be to validate these drugs in animal models of TNBC.

In conclusion, I have demonstrated the utility of using both VS and HTS methods to search for novel drug inhibitors of the RSK protein. Comparing the two lists suggested potential false positives and false negatives as well as highlighted the strengths and weaknesses of each method. These results agree with previous studies in that VS and HTS are parallel approaches producing two different, but useful, lists of inhibitors [62-65]. The improved docking pipeline developed here can be used to screen future chemical libraries consisting of more approved drugs, novel inhibitors, or derivative compounds of existing known inhibitors. Three of the four drugs determined to be novel RSK inhibitors are nutritional supplements with known safety profiles. The fourth drug is already a cancer agent with a known ability to modulate TP53 activity. Thus, these drugs are prime candidates to be repurposed to RSK-related cancers.

6.4 Finding personalized drug options for a patient with a rare tumor

The cost of sequencing the human genome is rapidly declining, and with it, whole genome sequencing for medical use is advancing. In Chapter 5, we were able to use whole genome and transcriptome sequencing to generate a hypothesis about the disease mechanisms of a patient with a rare tongue tumor and consequently reposition renal cancer drugs to treat this tumor. This was one of the first personalized medicine studies using sequencing and highlights the utility of genomic data to inform rational therapeutic options.

There were many advantages to a sequencing approach. First, sequencing can be performed with very small samples, such as from a fine needle biopsy. Compared to clinical genetic testing for patients with risk of disease, genotyping or genomic assays, the sequencing approach is comprehensive and can determine copy number, expression, and sequence changes concurrently. It is not limited to known SNPs and risk factors, or a small subset of genes. Being able to study the entire genome allowed us to form a hypothesis about the disease mechanism of the tumor. For example, PTEN is most often associated with upregulation of the PI3K/AKT pathway; however, we observed apparent downregulation of this pathway in the tumor and thus focused on the RET-signaling pathway instead. When we found potential therapeutic options, I was able to confirm that the known targets of the proposed drugs were expressed and did not have any interfering mutations. This type of analysis can be expanded to include drug metabolizing enzymes, for example, and would add insight into whether the drug would be effective for the patient.

Genome sequencing of the metastasized skin tumor also revealed several important observations. First, there were extensive large-scale DNA changes compared to the original tumor, highlighting the genomic instability of the post-treatment metastasis. Second, many of the changes appeared to correspond with the tumor acquiring resistance to RET-targeted therapies. Third, knowing the genetic makeup of the tumor allowed us to suggest a cocktail of rational targeted therapies for the patient; however, determining the proper dosages and potential drug-drug interactions in a large drug cocktail would be a difficult task, requiring further *in vitro* and *in vivo* research.

There were also limitations to our study. The significant genome contamination of the lung tumor transcriptome with genomic DNA may have affected the gene expression analysis. However, after correcting for background contamination, I believe that the most important signals should still be retained, just at a lower level of coverage. Indeed, we were still able to detect the highly amplified genes and showed that they correlated well with copy number data for both the lung and skin tumor genomes (Appendix D). Adjacent normal tissue for the tumors was unavailable for differential expression analysis. Instead tumor expression was compared to blood expression and the compendium. Expression differences between tumor and blood may have been caused by inherent differences between the tissue types. By including the compendium, these types of differences could be eliminated. However, the composition of the compendium is also significant; for example, if many cancers in the compendium were driven by RET, we would not have been able to detect RET as important in this patient. Lastly, we could not be sure that the skin tumor metastasized from the lung tumor. However, the presence of the four somatic mutations in both the lung and skin tumors suggests that we could meaningfully compare the skin and lung tumors as post- and pre-treatment tumors.

Our study highlighted key challenges in cancer drug discovery. For instance, there were many aberrations in the cancer genomes, which played important roles in driving the disease but were not targetable by known drugs, such as the down-regulation or mutation of the tumor suppressor genes PTEN, TP53, and RB1. Existing drug discovery methods search for small molecule drug inhibitors of target proteins, but it is extremely challenging to develop novel small molecule agonists of mutated or downregulated targets. Other methods like siRNA, immunotherapy, and gene therapy methods have not yet progressed to widespread use. A second was the acquiring of resistance, for which there are two reasonable causes. First, the standard dosage of sunitinib was reduced after one round due to adverse effects. This is compounded by the second reason, which is that overexpression of the RET protein may actually have required higher doses of sunitinib in order to inhibit MAPK pathway signaling. Another challenge is the lack of safe and effective drugs for many proteins. In the skin metastasis, drugs that inhibit AKT, the mTOR-RICTOR, MEK1, or ERK1 would have been valuable for this patient. However, no approved drugs for these targets exist. In addition,

aside from AKT, these proteins have not yet been proven to drive proliferation in specific cancers, and as such it would be difficult to identify a patient subset to test such inhibitors in clinical trials.

What this study did show is that whole genome and transcriptome sequencing of patient genomes is now tractable and can both inform disease mechanisms and rational therapeutic options. This approach is especially relevant for rare tumors with poorly understand molecular mechanisms and no standard treatment options since the scarcity of these patients and the diversity of their tumors would be challenging for establishing clinical trials. It is thus conceivable that characterizing tumor genomes will become a routine part of cancer therapy in the future, with continued monitoring to update treatment options.

6.5 Conclusion

Drug candidates with low efficacy but good safety profiles may fail clinical trials due to poor target inhibition in humans. In such a case, my molecular docking approach can help determine novel therapeutic targets for this drug. Candidates may also fail due to heterogeneity of the disease across patients. The sequencing approach can help elucidate the disease mechanism for each patient, and find the subset for which the drug would be effective. I would thus be able to inform therapeutic options and conduct clinical trials for patients on a molecularly-personalized level. On the whole, the research presented in this thesis provides new methods to improve drug discovery efficiency.

Bibliography

1. Zoete V, Grosdidier A, Michielin O. (2009) Docking, virtual high throughput screening and in silico fragment-based drug design. J Cell Mol Med 13: 238-248. 10.1111/j.1582-4934.2008.00665.x.

2. Kitchen DB, Decornez H, Furr JR, Bajorath J. (2004) Docking and scoring in virtual screening for drug discovery: Methods and applications. Nat Rev Drug Discov 3: 935-949. 10.1038/nrd1549. 12/10/2007.

3. Wu KK. (2000) Aspirin and salicylate: An old remedy with a new twist Circulation 102: 2022-2023.

4. Henderson JW. (1997) The yellow brick road to penicillin: A story of serendipity Mayo Clin Proc 72: 683-687.

5. Chin YW, Balunas MJ, Chai HB, Kinghorn AD. (2006) Drug discovery from natural sources. AAPS J 8: E239-53. 10.1208/aapsj080228.

6. Mueller RL, Scheidt S. (1994) History of drugs for thrombotic disease. discovery, development, and directions for the future. Circulation 89: 432-449.

7. Drews J. (2000) Drug discovery: A historical perspective Science 287: 1960-1964.

8. Gibbs JB. (2000) Mechanism-based target identification and drug discovery in cancer research Science 287: 1969 <last_page> 1973. 10.1126/science.287.5460.1969.

9. Simmons KJ, Chopra I, Fishwick CWG. (2010) Structure-based discovery of antibacterial drugs Nature Reviews Microbiology 8: 501-510. 10.1038/nrmicro2349. Available: http://www.nature.com/nrmicro/current_issue/ via the Internet.

10. Munos B. (2009) Lessons from 60 years of pharmaceutical innovation Nat Rev Drug Discov 8: 959-968. 10.1038/nrd2961. Available: www.refworks.com via the Internet.

11. Lawrence S. (2007) Drug output slows in 2006. Nat Biotechnol 25: 1073. 10.1038/nbt1007-1073. 12/10/2007.

12. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, *et al*. (2010) How to improve R&D productivity: The pharmaceutical industry's grand challenge Nature Reviews Drug Discovery . 10.1038/nrd3078. Available: http://www.nature.com/nrd/current_issue/ via the Internet.

13. Kola I. (2008) The state of innovation in drug development. Clin Pharmacol Ther 83: 227-230. 10.1038/sj.clpt.6100479.

14. Kola I, Landis J. (2004) Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3: 711-715. 10.1038/nrd1470.

15. Ashburn TT, Thor KB. (2004) Drug repositioning: Identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3: 673-683. 10.1038/nrd1468. 12/10/2007.

16. Thomas CC, Deak M, Alessi DR, van Aalten DM. (2002) High-resolution structure of the pleckstrin homology domain of protein kinase b/akt bound to phosphatidylinositol (3,4,5)-trisphosphate. Curr Biol 12: 1256-1262.

17. Choi YH, Yoon CJ, Park JH, Chung JW, Kwon JW, *et al*. (2003) Balloon-occluded retrograde transvenous obliteration for gastric variceal bleeding: Its feasibility compared with transjugular intrahepatic portosystemic shunt. Korean J Radiol 4: 109-116.

18. Bradley D. (2005) Why big pharma needs to learn the three 'R's. Nat Rev Drug Discov 4: 446. 12/10/2007.

19. Druker B. (2004) Advances in cancer research volume 91; imatinib as a paradigm of targeted therapies 91: 1 <last_page> 30. 10.1016/S0065-230X(04)91001-9.

20. Imming P, Sinning C, Meyer A. (2006) Drugs, their targets and the nature and number of drug targets Nat Rev Drug Discov 5: 821-834. 10.1038/nrd2132.

21. Imai K, Takaoka A. (2006) Comparing antibody and small-molecule therapies for cancer. Nat Rev Cancer 6: 714-727. 10.1038/nrc1913.

22. Chames P, Van Regenmortel M, Weiss E, Baty D. (2009) Therapeutic antibodies: Successes, limitations and hopes for the future. Br J Pharmacol 157: 220-233. 10.1111/j.1476-5381.2009.00190.x.

23. Arkin MR, Wells JA. (2004) Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. Nat Rev Drug Discov 3: 301-317. 10.1038/nrd1343.

24. Fletcher S, Hamilton AD. (2005) Protein surface recognition and proteomimetics: Mimics of protein surface structure and function. Curr Opin Chem Biol 9: 632-638. 10.1016/j.cbpa.2005.10.006.

25. Hopkins AL, Groom CR. (2002) The druggable genome. Nat Rev Drug Discov 1: 727-730. 10.1038/nrd892.

26. Overington JP, Al-Lazikani B, Hopkins AL. (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993-996. 10.1038/nrd2199.

27. Knight ZA, Lin H, Shokat KM. (2010) Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 10: 130-137. 10.1038/nrc2787.

28. Lazo JS. (2008) Rear-view mirrors and crystal balls: A brief reflection on drug discovery. Mol Interv 8: 60-63. 10.1124/mi.8.2.1.

29. Vane JR, Botting RM. (2003) The mechanism of action of aspirin. Thromb Res 110: 255-258.

30. Emmerson AM, Jones AM. (2003) The quinolones: Decades of development and use. J Antimicrob Chemother 51 Suppl 1: 13-20. 10.1093/jac/dkg208.

31. Drlica K, Zhao X. (1997) DNA gyrase, topoisomerase IV, and the 4-quinolones. Microbiol Mol Biol Rev 61: 377-392.

32. Lokey RS. (2003) Forward chemical genetics: Progress and obstacles on the path to a new pharmacopoeia Curr Opin Chem Biol 7: 91-96.

33. Inglese J, Johnson RL, Simeonov A, Xia M, Zheng W, *et al*. (2007) High-throughput screening assays for the identification of chemical probes. Nat Chem Biol 3: 466-479. 10.1038/nchembio.2007.17.

34. Strebhardt K, Ullrich A. (2008) Paul ehrlich's magic bullet concept: 100 years of progress. Nat Rev Cancer 8: 473-480. 10.1038/nrc2394.

35. Houshmand P, Zlotnik A. (2003) Targeting tumor cells. Curr Opin Cell Biol 15: 640-644.

36. Goldman JM, Melo JV. (2001) Targeting the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 344: 1084-1086. 10.1056/NEJM200104053441409.

37. Ghosh S, Collins FS. (1996) The geneticist's approach to complex disease. Annu Rev Med 47: 333-353. 10.1146/annurev.med.47.1.333.

38. Marusyk A, Polyak K. (2010) Tumor heterogeneity: Causes and consequences. Biochim Biophys Acta 1805: 105-117. 10.1016/j.bbcan.2009.11.002.

39. De Clercq E. (2009) Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. Int J Antimicrob Agents 33: 307-320. 10.1016/j.ijantimicag.2008.10.010.

40. Roth BL, Sheffler DJ, Kroeze WK. (2004) Magic shotguns versus magic bullets: Selectively non-selective drugs for mood disorders and schizophrenia. Nat Rev Drug Discov 3: 353-359. 10.1038/nrd1346.

41. Potapova O, Laird AD, Nannini MA, Barone A, Li G, *et al*. (2006) Contribution of individual targets to the antitumor efficacy of the multitargeted receptor tyrosine kinase inhibitor SU11248. Mol Cancer Ther 5: 1280-1289. 10.1158/1535-7163.MCT-03-0156.

42. Fabian MA, Biggs WH,3rd, Treiber DK, Atteridge CE, Azimioara MD, *et al*. (2005) A small molecule-kinase interaction map for clinical kinase inhibitors. Nat Biotechnol 23: 329-336. 10.1038/nbt1068.

43. Brehmer D, Greff Z, Godl K, Blencke S, Kurtenbach A, *et al*. (2005) Cellular targets of gefitinib. Cancer Res 65: 379-382.

44. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, *et al*. (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: New pharmacological opportunities due to related binding site recognition. J Med Chem 47: 550-557. 10.1021/jm030912m.

45. Maier TJ, Tausch L, Hoernig M, Coste O, Schmidt R, *et al*. (2008) Celecoxib inhibits 5-lipoxygenase. Biochem Pharmacol 76: 862-872. 10.1016/j.bcp.2008.07.009.

46. Claria J, Romano M. (2005) Pharmacological intervention of cyclooxygenase-2 and 5-lipoxygenase pathways. impact on inflammation and cancer. Curr Pharm Des 11: 3431-3447.

47. Du L, Li M, You Q, Xia L. (2007) A novel structure-based virtual screening model for the hERG channel blockers. Biochem Biophys Res Commun 355: 889-894. 10.1016/j.bbrc.2007.02.068.

48. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL. (2006) Global mapping of pharmacological space Nat Biotechnol 24: 805-815. 10.1038/nbt1228.

49. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. (2007) Drug-target network Nat Biotechnol 25: 1119-1126. 10.1038/nbt1338.

50. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV. (2009) The topology of drug-target interaction networks: Implicit dependence on drug properties and target families. Mol Biosyst 5: 1051-1057. 10.1039/b905821b.

51. Karaman MW, Herrgard S, Treiber DK, Gallant P, Atteridge CE, *et al*. (2008) A quantitative analysis of kinase inhibitor selectivity. Nat Biotechnol 26: 127-132. 10.1038/nbt1358.

52. Jorgensen WL. (2004) The many roles of computation in drug discovery. Science 303: 1813-1818. 10.1126/science.1096361. 12/10/2007.

53. Kubinyi H. (1997) QSAR and 3D QSAR in drug design part 1: Methodology Drug Discov Today 2: 457 <last_page> 467. 10.1016/S1359-6446(97)01079-9.

54. Schneider G, Fechner U. (2005) Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov 4: 649-663. 10.1038/nrd1799.

55. Parker CN, Bajorath J. (2006) Towards unified compound screening strategies: A critical evaluation of error sources in experimental and virtual high-throughput screening. QSAR & Combinatorial Science 25: 11531161.

56. Shoichet BK. (2006) Screening in a spirit haunted world. Drug Discov Today 11: 607-615. 10.1016/j.drudis.2006.05.014.

57. Crisman TJ, Parker CN, Jenkins JL, Scheiber J, Thoma M, *et al*. (2007) Understanding false positives in reporter gene assays: In silico chemogenomics approaches to prioritize cell-based HTS data. J Chem Inf Model 47: 1319-1327. 10.1021/ci6005504.

58. Feng BY, Simeonov A, Jadhav A, Babaoglu K, Inglese J, *et al*. (2007) A high-throughput screen for aggregation-based inhibition in a large compound library. J Med Chem 50: 2385-2390. 10.1021/jm061317y.

59. Ferreira RS, Simeonov A, Jadhav A, Eidam O, Mott BT, *et al*. (2010) Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. J Med Chem 53: 4891-4905. 10.1021/jm100488w.

60. Li JW, Vederas JC. (2009) Drug discovery and natural products: End of an era or an endless frontier? Science 325: 161-165. 10.1126/science.1168243.

61. Mishra KP, Ganju L, Sairam M, Banerjee PK, Sawhney RC. (2008) A review of high throughput technology for the screening of natural products. Biomed Pharmacother 62: 94-98. 10.1016/j.biopha.2007.06.012.

62. Edwards BS, Bologa C, Young SM, Balakin KV, Prossnitz ER, *et al*. (2005) Integration of virtual screening with high-throughput flow cytometry to identify novel small molecule formylpeptide receptor antagonists. Mol Pharmacol 68: 1301-1310. 10.1124/mol.105.014068.

63. Polgar T, Baki A, Szendrei GI, Keseru GM. (2005) Comparative virtual and experimental high-throughput screening for glycogen synthase kinase-3beta inhibitors. J Med Chem 48: 7946-7959. 10.1021/jm050504d.

64. Paiva AM, Vanderwall DE, Blanchard JS, Kozarich JW, Williamson JM, *et al*. (2001) Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of mycobacterium tuberculosis. Biochim Biophys Acta 1545: 67-77.

65. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, *et al*. (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45: 2213-2221.

66. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK. (2009) Docking and chemoinformatics screens for new ligands and targets. Curr Opin Biotechnol 20: 429-436.

67. [Anonymous]. RCSB PDB statistics. 2011.

68. Acharya KR, Lloyd MD. (2005) The advantages and limitations of protein crystal structures. Trends Pharmacol Sci 26: 10-14. 10.1016/j.tips.2004.10.011.

69. Jenny Gu, Bourne PE. (2009) Structural bioinformatics Hoboken, N.J.: Wiley-Blackwell. 1035 p.

70. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci U S A 104: 9615-9620. 10.1073/pnas.0610313104.

71. Spronk CA, Linge JP, Hilbers CW, Vuister GW. (2002) Improving the quality of protein structures derived by NMR spectroscopy. J Biomol NMR 22: 281-289.

72. Ginalski K. (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16: 172-177. 10.1016/j.sbi.2006.02.003.

73. McGovern SL, Shoichet BK. (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 46: 2895-2907. 10.1021/jm0300330.

74. Evers A, Klebe G. (2004) Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. J Med Chem 47: 5381-5392. 10.1021/jm0311487.

75. Kairys V, Fernandes MX, Gilson MK. (2006) Screening drug-like compounds by docking to homology models: A systematic study. J Chem Inf Model 46: 365-379. 10.1021/ci050238c.

76. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, *et al*. (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58: 899-907.

77. Laurie AT, Jackson RM. (2005) Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21: 1908-1916. 10.1093/bioinformatics/bti315.

78. An J, Totrov M, Abagyan R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 4: 752-761.

10.1074/mcp.M400159-MCP200. 12/10/2007.

79. Brady GP,Jr, Stouten PF. (2000) Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 14: 383-401.

80. Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, *et al*. (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science 289: 1938-1942.

81. Zhang J, Adrian FJ, Jahnke W, Cowan-Jacob SW, Li AG, *et al*. (2010) Targeting bcr-abl by combining allosteric with ATP-binding-site inhibitors. Nature 463: 501-506. 10.1038/nature08675.

82. Adrian FJ, Ding Q, Sim T, Velentza A, Sloan C, *et al*. (2006) Allosteric inhibitors of bcr-abl-dependent cell proliferation. Nat Chem Biol 2: 95-102. 10.1038/nchembio760.

83. Jain AN. (2003) Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 46: 499-511. 10.1021/jm020406h.

84. Ewing TJ, Makino S, Skillman AG, Kuntz ID. (2001) DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15: 411-428.

85. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46: 3-26.

86. Lengauer T, Rarey M. (1996) Computational methods for biomolecular docking. Curr Opin Struct Biol 6: 402-406.

87. Abagyan R, Totrov M, Kuznetsov D. (1994) ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. Journal of Computational Chemistry 15: 488. 10.1002/jcc.540150503. 12/10/2007.

88. Abagyan R, Totrov M. (1994) Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 235: 983-1002. 10.1006/jmbi.1994.1052.

89. Totrov M, Abagyan R. (1994) Efficient parallelization of the energy, surface, and derivative calculations for internal coordinate mechanics Journal of Computational Chemistry 15: 1105 <last_page> 1112. 10.1002/jcc.540151006.

90. Abagyan R, Orry A, Raush E, Budagyan L, Totrov M. (2007) ICM manual. version 3.0.

91. Luo W, Pei J, Zhu Y. (2010) A fast protein-ligand docking algorithm based on hydrogen bond matching and surface shape complementarity. J Mol Model 16: 903-913.

10.1007/s00894-009-0598-7.

92. Fuhrmann J, Rurainski A, Lenhof HP, Neumann D. (2010) A new lamarckian genetic algorithm for flexible ligand-receptor docking. J Comput Chem 31: 1911-1918. 10.1002/jcc.21478.

93. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. Br J Pharmacol 153 Suppl 1: S7-26. 10.1038/sj.bjp.0707515.

94. Liebeschuetz JW. (2008) Evaluating docking programs: Keeping the playing field level. J Comput Aided Mol Des 22: 229-238. 10.1007/s10822-008-9169-8.

95. Li X, Li Y, Cheng T, Liu Z, Wang R. (2010) Evaluation of the performance of fmy molecular docking programs on a diverse set of protein-ligand complexes. J Comput Chem 31: 2109-2125. 10.1002/jcc.21498.

96. Perola E, Walters WP, Charifson PS. (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56: 235-249. 10.1002/prot.20088.

97. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J. (2006) On evaluating molecular-docking methods for pose prediction and enrichment factors. J Chem Inf Model 46: 401-415. 10.1021/ci0503255.

98. Bursulaya BD, Totrov M, Abagyan R, Brooks CL,3rd. (2003) Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 17: 755-763.

99. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, *et al*. (2009) Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. J Chem Inf Model 49: 1455-1474. 10.1021/ci900056c.

100. Huang N, Shoichet BK, Irwin JJ. (2006) Benchmarking sets for molecular docking. J Med Chem 49: 6789-6801. 10.1021/jm0608356.

101. Rich RL, Day YS, Morton TA, Myszka DG. (2001) High-resolution and high-throughput protocols for measuring drug/human serum albumin interactions using BIACORE. Anal Biochem 296: 197-207. 10.1006/abio.2001.5314.

102. Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, *et al*. (2001) High-density miniaturized thermal shift assays as a general strategy for drug discovery. J Biomol Screen 6: 429-440. 10.1089/108705701753364922.

103. Lomenick B, Hao R, Jonai N, Chin RM, Aghajan M, *et al*. (2009) Target identification using drug affinity responsive target stability (DARTS). Proc Natl Acad Sci U S A 106:

21984-21989. 10.1073/pnas.0910040106.

104. von Rechenberg M, Blake BK, Ho YS, Zhen Y, Chepanoske CL, *et al*. (2005) Ampicillin/penicillin-binding protein interactions as a model drug-target system to optimize affinity pull-down and mass spectrometric strategies for target and pathway identification. Proteomics 5: 1764-1773. 10.1002/pmic.200301088.

105. Oprea T, Tropsha A. (2006) Target, chemical and bioactivity databases – integration is key. Drug Discovery Today: Technologies 3: 357 <last_page> 365. 10.1016/j.ddtec.2006.12.003.

106. Klebe G. (2006) Virtual ligand screening: Strategies, perspectives and limitations. Drug Discov Today 11: 580-594. 10.1016/j.drudis.2006.05.012.

107. DePristo MA, de Bakker PI, Blundell TL. (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 12: 831-838. 10.1016/j.str.2004.02.031.

108. Murray CW, Baxter CA, Frenkel AD. (1999) The sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. J Comput Aided Mol Des 13: 547-562.

109. Plewczynski D, Lazniewski M, Augustyniak R, Ginalski K. (2011) Can I trust docking results? evaluation of seven commonly used programs on PDBbind database. J Comput Chem 32: 742-755. 10.1002/jcc.21643; 10.1002/jcc.21643.

110. Wermuth C. (2006) Selective optimization of side activities: The SOSA approach Drug Discov Today 11: 160 <last_page> 164. 10.1016/S1359-6446(05)03686-X.

111. Chong CR, Sullivan DJ,Jr. (2007) New uses for old drugs. Nature 448: 645-646. 10.1038/448645a.

112. Borisy AA, Elliott PJ, Hurst NW, Lee MS, Lehar J, *et al*. (2003) Systematic discovery of multicomponent therapeutics. Proc Natl Acad Sci U S A 100: 7977-7982. 10.1073/pnas.1337088100.

113. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, *et al*. (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science 313: 1929-1935. 10.1126/science.1132939.

114. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, *et al*. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A 107: 14621-14626. 10.1073/pnas.1000138107.

115. Chen YZ, Ung CY. (2001) Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. J Mol Graph Model

20: 199-218.

116. Chen YZ, Zhi DG. (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 43: 217-226. 12/10/2007.

117. Zahler S, Tietze S, Totzke F, Kubbutat M, Meijer L, *et al*. (2007) Inverse in silico screening for identification of kinase inhibitor targets. Chem Biol 14: 1207-1214. 10.1016/j.chembiol.2007.10.010.

118. Li H, Gao Z, Kang L, Zhang H, Yang K, *et al*. (2006) TarFisDock: A web server for identifying drug targets with docking approach. Nucleic Acids Res 34: W219-24. 10.1093/nar/gkl114. 12/10/2007.

119. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. (2008) Drug target identification using side-effect similarity. Science 321: 263-266. 10.1126/science.1158140. 8/1/2008.

120. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, *et al*. (2009) Predicting new molecular targets for known drugs. Nature 462: 175-181. 10.1038/nature08506.

121. Durrant JD, Amaro RE, Xie L, Urbaniak MD, Ferguson MA, *et al*. (2010) A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. PLoS Comput Biol 6: e1000648. 10.1371/journal.pcbi.1000648.

122. Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, *et al*. (2009) Drug discovery using chemical systems biology: Repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol 5: e1000423. 10.1371/journal.pcbi.1000423.

123. Ha S, Seo YJ, Kwon MS, Chang BH, Han CK, *et al*. (2008) IDMap: Facilitating the detection of potential leads with therapeutic targets. Bioinformatics 24: 1413-1415. 10.1093/bioinformatics/btn138. 8/1/2008.

124. Sotriffer CA, Dramburg I. (2005) "In situ cross-docking" to simultaneously address multiple targets. J Med Chem 48: 3122-3125. 10.1021/jm050075j.

125. Li L, Bum-Erdene K, Baenziger PH, Rosen JJ, Hemmert JR, *et al*. (2010) BioDrugScreen: A computational drug design resource for ranking molecules docked to the human proteome. Nucleic Acids Res 38: D765-73. 10.1093/nar/gkp852.

126. Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, *et al*. (2006) Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics 7: 270. 10.1186/1471-2105-7-270.

127. Cherkasov A, Shi Z, Li Y, Jones SJ, Fallahi M, *et al*. (2005) 'Inductive' charges on atoms in proteins: Comparative docking with the extended steroid benchmark set and discovery of a novel SHBG ligand. J Chem Inf Model 45: 1842-1853. 10.1021/ci0498158.

128. Cherkasov A, Ban F, Li Y, Fallahi M, Hammond GL. (2006) Progressive docking: A hybrid QSAR/docking approach for accelerating in silico high throughput screening. J Med Chem 49: 7466-7478. 10.1021/jm060961+.

129. Li YY, Jones SJ, Cherkasov A. (2006) Selective targeting of indel-inferred differences in spatial structures of homologous proteins. J Bioinform Comput Biol 4: 403-414.

130. Nandan D, Lopez M, Ban F, Huang M, Li Y, *et al*. (2007) Indel-based targeting of essential proteins in human pathogens that have close host orthologue(s): Discovery of selective inhibitors for leishmania donovani elongation factor-1alpha. Proteins 67: 53-64. 10.1002/prot.21278.

131. Yakovenko OY, Li YY, Oliferenko AA, Vashchenko GM, Bdzhola VG, *et al*. (2011) Ab initio parameterization of YFF1, a universal force field for drug-design applications. J Mol Model . 10.1007/s00894-011-1095-3.

132. Goodsell DS, Morris GM, Olson AJ. (1996) Automated docking of flexible ligands: Applications of AutoDock. J Mol Recognit 9: 1-5. 2-6.

133. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP. (2007) eHiTS: A new fast, exhaustive flexible ligand docking system. J Mol Graph Model 26: 198-212. 10.1016/j.jmgm.2006.06.002.

134. Kramer B, Rarey M, Lengauer T. (1999) Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 37: 228-241.

135. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. (2003) Improved protein-ligand docking using GOLD. Proteins 52: 609-623. 10.1002/prot.10465.

136. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, *et al*. (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. J Med Chem 47: 1739-1749. 10.1021/jm0306430.

137. [Anonymous]. The NCI/DTP open chemical repository. .

138. Irwin JJ, Shoichet BK. (2005) ZINC--a free database of commercially available compounds for virtual screening. J Chem Inf Model 45: 177-182. 10.1021/ci049714+.

139. Chen J, Swamidass SJ, Dou Y, Bruand J, Baldi P. (2005) ChemDB: A public database of small molecules and related chemoinformatics resources. Bioinformatics 21: 4133-4139. 10.1093/bioinformatics/bti683.

140. Li Q, Cheng T, Wang Y, Bryant SH. (2010) PubChem as a public resource for drug discovery. Drug Discov Today 15: 1052-1057. 10.1016/j.drudis.2010.10.003.

141. Blum LC, Reymond JL. (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc 131: 8732-8733. 10.1021/ja902302h.

142. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, *et al*. (2009) PubChem: A public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37: W623-33. 10.1093/nar/gkp456.

143. Han L, Suzek TO, Wang Y, Bryant SH. (2010) The text-mining based PubChem bioassay neighboring analysis. BMC Bioinformatics 11: 549. 10.1186/1471-2105-11-549.

144. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, *et al*. (2006) DrugBank: A comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34: D668-72. 10.1093/nar/gkj067. 12/10/2007.

145. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, *et al*. (2006) From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res 34: D354-7. 10.1093/nar/gkj102. 8/1/2008.

146. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, *et al*. (2005) WOMBAT: World of molecular bioactivity. In: Oprea TI, editor. Chemoinformatics in Drug Discovery. : Wiley-VCH. pp. 223-223–239.

147. Molecular Design, Ltd., San Leandro, CA. MDDR. .

148. Sheridan RP, Shpungin J. (2004) Calculating similarities between biological activities in the MDL drug data report database. J Chem Inf Comput Sci 44: 727-740. 10.1021/ci034245h.

149. von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, *et al*. (2011) PROMISCUOUS: A database for network-based drug-repositioning. Nucleic Acids Res 39: D1060-6. 10.1093/nar/gkq1037.

150. Taboureau O, Nielsen SK, Audouze K, Weinhold N, Edsgard D, *et al*. (2011) ChemProt: A disease chemical biology database. Nucleic Acids Res 39: D367-72. 10.1093/nar/gkq906.

151. Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, *et al*. (2010) Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin Nat Genet . 10.1038/ng.518. Available: www.refworks.com via the Internet.

152. Shah SP, Kobel M, Senz J, Morin RD, Clarke BA, *et al*. (2009) Mutation of FOXL2 in granulosa-cell tumors of the ovary N Engl J Med 360: 2719-2729. 10.1056/NEJMoa0902542.

153. Xie L, Li J, Xie L, Bourne PE. (2009) Drug discovery using chemical systems biology: Identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. PLoS Comput Biol 5: e1000387. 10.1371/journal.pcbi.1000387.

154. Brylinski M, Skolnick J. (2010) Cross-reactivity virtual profiling of the human kinome by X-react(KIN): A chemical systems biology approach. Mol Pharm 7: 2324-2333. 10.1021/mp1002976.

155. Rockey WM, Elcock AH. (2002) Progress toward virtual screening for drug side effects. Proteins 48: 664-671. 10.1002/prot.10186.

156. Yang L, Chen J, Shi L, Hudock MP, Wang K, *et al*. (2010) Identifying unexpected therapeutic targets via chemical-protein interactome. PLoS One 5: e9568. 10.1371/journal.pone.0009568.

157. Abagyan R, Totrov M. (2001) High-throughput docking for lead generation. Curr Opin Chem Biol 5: 375-382.

158. Cavasotto CN, Ortiz MA, Abagyan RA, Piedrafita FJ. (2006) In silico identification of novel EGFR inhibitors with antiproliferative activity against cancer cells. Bioorg Med Chem Lett 16: 1969-1974. 10.1016/j.bmcl.2005.12.067.

159. Jain AN. (2009) Effects of protein conformation in docking: Improved pose prediction through protein pocket adaptation. J Comput Aided Mol Des 23: 355-374. 10.1007/s10822-009-9266-3.

160. Kumar S, Boehm J, Lee JC. (2003) p38 MAP kinases: Key signalling molecules as therapeutic targets for inflammatory diseases. Nat Rev Drug Discov 2: 717-726. 10.1038/nrd1177.

161. Verdonk ML, Mortenson PN, Hall RJ, Hartshorn MJ, Murray CW. (2008) Protein-ligand docking against non-native protein conformers. J Chem Inf Model 48: 2214-2225. 10.1021/ci8002254.

162. Rix U, Hantschel O, Durnberger G, Remsing Rix LL, Planyavsky M, *et al*. (2007) Chemical proteomic profiles of the BCR-ABL inhibitors imatinib, nilotinib, and dasatinib reveal novel kinase and nonkinase targets. Blood 110: 4055-4063. 10.1182/blood-2007-07-102061.

163. Namboodiri HV, Bukhtiyarova M, Ramcharan J, Karpusas M, Lee Y, *et al*. (2010) Analysis of imatinib and sorafenib binding to p38alpha compared with c-abl and b-raf provides structural insights for understanding the selectivity of inhibitors targeting the

DFG-out form of protein kinases. Biochemistry 49: 3611-3618. 10.1021/bi100070r.

164. Manley PW, Drueckes P, Fendrich G, Furet P, Liebetanz J, *et al*. (2010) Extended kinase profile and properties of the protein kinase inhibitor nilotinib. Biochim Biophys Acta 1804: 445-453. 10.1016/j.bbapap.2009.11.008.

165. Koyama K, Hatsushika K, Ando T, Sakuma M, Wako M, *et al*. (2007) Imatinib mesylate both prevents and treats the arthritis induced by type II collagen antibody in mice. Mod Rheumatol 17: 306-310. 10.1007/s10165-007-0592-9.

166. Eklund KK, Lindstedt K, Sandler C, Kovanen PT, Laasonen L, *et al*. (2008) Maintained efficacy of the tyrosine kinase inhibitor imatinib mesylate in a patient with rheumatoid arthritis. J Clin Rheumatol 14: 294-296. 10.1097/RHU.0b013e318188b1ce.

167. Vernon MR, Pearson L, Atallah E. (2009) Resolution of rheumatoid arthritis symptoms with imatinib mesylate. J Clin Rheumatol 15: 267. 10.1097/RHU.0b013e3181b0d352.

168. Akashi N, Matsumoto I, Tanaka Y, Inoue A, Yamamoto K, *et al*. (2010) Comparative suppressive effects of tyrosine kinase inhibitors imatinib and nilotinib in models of autoimmune arthritis. Mod Rheumatol . 10.1007/s10165-010-0392-5.

169. Davies SP, Reddy H, Caivano M, Cohen P. (2000) Specificity and mechanism of action of some commonly used protein kinase inhibitors. Biochem J 351: 95-105.

170. Komander D, Kular GS, Schuttelkopf AW, Deak M, Prakash KR, *et al*. (2004) Interactions of LY333531 and other bisindolyl maleimide inhibitors with PDK1. Structure 12: 215-226. 10.1016/j.str.2004.01.005.

171. Fedorov O, Marsden B, Pogacic V, Rellos P, Muller S, *et al*. (2007) A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. Proc Natl Acad Sci U S A 104: 20523-20528. 10.1073/pnas.0708800104.

172. Cotreau MM, Stonis L, Dykstra KH, Gandhi T, Gutierrez M, *et al*. (2002) Multiple-dose, safety, pharmacokinetics, and pharmacodynamics of a new selective estrogen receptor modulator, ERA-923, in healthy postmenopausal women. J Clin Pharmacol 42: 157-165.

173. Kuiper GG, Carlsson B, Grandien K, Enmark E, Haggblad J, *et al*. (1997) Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. Endocrinology 138: 863-870.

174. Naldi L, Raho G. (2009) Emerging drugs for psoriasis. Expert Opin Emerg Drugs 14: 145-163. 10.1517/14728210902771334.

175. Pinette KV, Yee YK, Amegadzie BY, Nagpal S. (2003) Vitamin D receptor as a drug discovery target. Mini Rev Med Chem 3: 193-204.

176. Fuhrmann U, Krattenmacher R, Slater EP, Fritzemeier KH. (1996) The novel progestin drospirenone and its natural counterpart progesterone: Biochemical profile and antiandrogenic potential. Contraception 54: 243-251.

177. Wood JM, Bold G, Buchdunger E, Cozens R, Ferrari S, *et al*. (2000) PTK787/ZK 222584, a novel and potent inhibitor of vascular endothelial growth factor receptor tyrosine kinases, impairs vascular endothelial growth factor-induced responses and tumor growth after oral administration. Cancer Res 60: 2178-2189.

178. Weisberg E, Manley PW, Breitenstein W, Bruggen J, Cowan-Jacob SW, *et al*. (2005) Characterization of AMN107, a selective inhibitor of native and mutant bcr-abl. Cancer Cell 7: 129-141. 10.1016/j.ccr.2005.01.007.

179. Balendiran GK, Schnutgen F, Scapin G, Borchers T, Xhong N, *et al*. (2000) Crystal structure and thermodynamic analysis of human brain fatty acid-binding protein. J Biol Chem 275: 27045-27054. 10.1074/jbc.M003001200.

180. Belayev L, Marcheselli VL, Khoutorova L, Rodriguez de Turco EB, Busto R, *et al*. (2005) Docosahexaenoic acid complexed to albumin elicits high-grade ischemic neuroprotection. Stroke 36: 118-123. 10.1161/01.STR.0000149620.74770.2e.

181. Bantscheff M, Eberhard D, Abraham Y, Bastuck S, Boesche M, *et al*. (2007) Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat Biotechnol 25: 1035-1044. 10.1038/nbt1328.

182. Ross-Macdonald P, de Silva H, Guo Q, Xiao H, Hung CY, *et al*. (2008) Identification of a nonkinase target mediating cytotoxicity of novel kinase inhibitors. Mol Cancer Ther 7: 3490-3498. 10.1158/1535-7163.MCT-08-0826.

183. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK. (2009) Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotechnol 20: 429-436. 10.1016/j.copbio.2009.08.003.

184. Sabio M, Jones K, Topiol S. (2008) Use of the X-ray structure of the beta2-adrenergic receptor for drug discovery. part 2: Identification of active compounds. Bioorg Med Chem Lett 18: 5391-5395. 10.1016/j.bmcl.2008.09.046.

185. Kinnings SL, Liu N, Tonge PJ, Jackson RM, Xie L, *et al*. (2011) A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J Chem Inf Model 51: 408-419. 10.1021/ci100369f.

186. Teramoto R, Fukunishi H. (2008) Consensus scoring with feature selection for structure-based virtual screening. J Chem Inf Model 48: 288-295. 10.1021/ci700239t.

187. Teramoto R, Fukunishi H. (2007) Supervised scoring models with docked ligand conformations for structure-based virtual screening. J Chem Inf Model 47: 1858-1867.

10.1021/ci700116z.

188. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365-370. 12/10/2007.

189. Raevsky OA, Trepalin SV, Trepalina HP, Gerasimenko VA, Raevskaja OE. (2002) SLIPPER-2001 -- software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity. J Chem Inf Comput Sci 42: 540-549.

190. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, *et al*. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504. 10.1101/gr.1239303.

191. UniProt Consortium. (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38: D142-8. 10.1093/nar/gkp846.

192. Fukuda M, Kojima T, Kabayama H, Mikoshiba K. (1996) Mutation of the pleckstrin homology domain of bruton's tyrosine kinase in immunodeficiency impaired inositol 1,3,4,5-tetrakisphosphate binding capacity. J Biol Chem 271: 30303-30306.

193. Bando S, Takano T, Yubisui T, Shirabe K, Takeshita M, *et al*. (2004) Structure of human erythrocyte NADH-cytochrome b5 reductase. Acta Crystallogr D Biol Crystallogr 60: 1929-1934. 10.1107/S0907444904020645.

194. Stankova J, Lawrance AK, Rozen R. (2008) Methylenetetrahydrofolate reductase (MTHFR): A novel target for cancer therapy. Curr Pharm Des 14: 1143-1150.

195. Greasley SE, Marsilje TH, Cai H, Baker S, Benkovic SJ, *et al*. (2001) Unexpected formation of an epoxide-derived multisubstrate adduct inhibitor on the active site of GAR transformylase. Biochemistry 40: 13538-13547.

196. Oelkers W. (2004) Drospirenone, a progestogen with antimineralocorticoid properties: A short review. Mol Cell Endocrinol 217: 255-261. 10.1016/j.mce.2003.10.030.

197. Yu K, Bayona W, Kallen CB, Harding HP, Ravera CP, *et al*. (1995) Differential activation of peroxisome proliferator-activated receptors by eicosanoids. J Biol Chem 270: 23975-23983.

198. Lebovitz H. (2006) Diabetes: Assessing the pipeline. Atheroscler Suppl 7: 43-49. 10.1016/j.atherosclerosissup.2006.01.007.

199. Klaholz BP, Mitschler A, Moras D. (2000) Structural basis for isotype selectivity of the human retinoic acid nuclear receptor. J Mol Biol 302: 155-170. 10.1006/jmbi.2000.4032.

200. Desai SH, Boskovic G, Eastham L, Dawson M, Niles RM. (2000) Effect of receptor-selective retinoids on growth and differentiation pathways in mouse melanoma cells. Biochem Pharmacol 59: 1265-1275.

201. Fanjul AN, Delia D, Pierotti MA, Rideout D, Yu JQ, *et al*. (1996) 4-hydroxyphenyl retinamide is a highly selective activator of retinoid receptors. J Biol Chem 271: 22441-22446.

202. Lengqvist J, Mata De Urquiza A, Bergman AC, Willson TM, Sjovall J, *et al*. (2004) Polyunsaturated fatty acids including docosahexaenoic and arachidonic acid bind to the retinoid X receptor alpha ligand-binding domain. Mol Cell Proteomics 3: 692-703. 10.1074/mcp.M400003-MCP200.

203. Normanno N, De Luca A, Bianco C, Strizzi L, Mancino M, *et al*. (2006) Epidermal growth factor receptor (EGFR) signaling in cancer. Gene 366: 2-16. 10.1016/j.gene.2005.10.018.

204. Ciardiello F, Tortora G. (2008) EGFR antagonists in cancer treatment. N Engl J Med 358: 1160-1174. 10.1056/NEJMra0707704.

205. Choowongkomon K, Sawatdichaikul O, Songtawee N, Limtrakul J. (2010) Receptor-based virtual screening of EGFR kinase inhibitors from the NCI diversity database. Molecules 15: 4041-4054. 10.3390/molecules15064041.

206. Li L, Khanna M, Jo I, Wang F, Ashpole NM, *et al*. (2011) Target-specific support vector machine scoring in structure-based virtual screening: Computational validation, in vitro testing in kinases, and effects on lung cancer cell proliferation. J Chem Inf Model 51: 755-759. 10.1021/ci100490w.

207. La Motta C, Sartini S, Tuccinardi T, Nerini E, Da Settimo F, *et al*. (2009) Computational studies of epidermal growth factor receptor: Docking reliability, three-dimensional quantitative structure-activity relationship analysis, and virtual screening studies. J Med Chem 52: 964-975. 10.1021/jm800829v.

208. Vieth M, Siegel MG, Higgs RE, Watson IA, Robertson DH, *et al*. (2004) Characteristic physical properties and structural fragments of marketed oral drugs. J Med Chem 47: 224-232. 10.1021/jm030267j.

209. Shewchuk L, Hassell A, Wisely B, Rocque W, Holmes W, *et al*. (2000) Binding mode of the 4-anilinoquinazoline class of protein kinase inhibitor: X-ray crystallographic studies of 4-anilinoquinazolines bound to cyclin-dependent kinase 2 and p38 kinase. J Med Chem 43: 133-138.

210. Liao Y, Zhao H, Ogai A, Kato H, Asakura M, *et al*. (2008) Atorvastatin slows the progression of cardiac remodeling in mice with pressure overload and inhibits epidermal

growth factor receptor activation. Hypertens Res 31: 335-344. 10.1291/hypres.31.335.

211. Honegger AM, Dull TJ, Felder S, Van Obberghen E, Bellot F, *et al*. (1987) Point mutation at the ATP binding site of EGF receptor abolishes protein-tyrosine kinase activity and alters cellular routing. Cell 51: 199-209.

212. Maceratesi P, Daude N, Dallapiccola B, Novelli G, Allen R, *et al*. (1998) Human UDP-galactose 4' epimerase (GALE) gene and identification of five missense mutations in patients with epimerase-deficiency galactosemia. Mol Genet Metab 63: 26-30. 10.1006/mgme.1997.2645.

213. Mathew G, Knaus SJ. (2006) Acquired fanconi's syndrome associated with tenofovir therapy. J Gen Intern Med 21: C3-5. 10.1111/j.1525-1497.2006.00518.x.

214. Gupta SK. (2008) Tenofovir-associated fanconi syndrome: Review of the FDA adverse event reporting system. AIDS Patient Care STDS 22: 99-103. 10.1089/apc.2007.0052.

215. Shimizu T, Ikegami T, Ogawara M, Suzuki Y, Takahashi M, *et al*. (2002) Transgenic expression of the protein-L-isoaspartyl methyltransferase (PIMT) gene in the brain rescues mice from the fatal epilepsy of PIMT deficiency. J Neurosci Res 69: 341-352. 10.1002/jnr.10301.

216. Stratford AL, Habibi G, Astanehe A, Jiang H, Hu K, *et al*. (2007) Epidermal growth factor receptor (EGFR) is transcriptionally induced by the Y-box binding protein-1 (YB-1) and can be inhibited with iressa in basal-like breast cancer, providing a potential target for therapy. Breast Cancer Res 9: R61. 10.1186/bcr1767.

217. Liao C, Sun Q, Liang B, Shen J, Shuai X. (2010) Targeting EGFR-overexpressing tumor cells using cetuximab-immunomicelles loaded with doxorubicin and superparamagnetic iron oxide. Eur J Radiol . 10.1016/j.ejrad.2010.08.005.

218. Lee-Hoeflich ST, Crocker L, Yao E, Pham T, Munroe X, *et al*. (2008) A central role for HER3 in HER2-amplified breast cancer: Implications for targeted therapy. Cancer Res 68: 5878-5887. 10.1158/0008-5472.CAN-08-0380.

219. Irwin ME, Mueller KL, Bohin N, Ge Y, Boerner JL. (2010) Lipid raft localization of EGFR alters the response of cancer cells to the EGFR tyrosine kinase inhibitor gefitinib. J Cell Physiol . 10.1002/jcp.22570.

220. Warburton C, Dragowska WH, Gelmon K, Chia S, Yan H, *et al*. (2004) Treatment of HER-2/neu overexpressing breast cancer xenograft models with trastuzumab (herceptin) and gefitinib (ZD1839): Drug combination effects on tumor growth, HER-2/neu and epidermal growth factor receptor expression, and viable hypoxic cell fraction. Clin Cancer Res 10: 2512-2524.

221. Ueda Y, Hirai S, Osada S, Suzuki A, Mizuno K, *et al*. (1996) Protein kinase C activates the MEK-ERK pathway in a manner independent of ras and dependent on raf. J Biol Chem 271: 23512-23519.

222. Kosugi S, Furuchi T, Katane M, Sekine M, Shirasawa T, *et al*. (2008) Suppression of protein l-isoaspartyl (d-aspartyl) methyltransferase results in hyperactivation of EGF-stimulated MEK-ERK signaling in cultured mammalian cells. Biochem Biophys Res Commun 371: 22-27. 10.1016/j.bbrc.2008.03.109.

223. Roberts PJ, Der CJ. (2007) Targeting the raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. Oncogene 26: 3291-3310. 10.1038/sj.onc.1210422.

224. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, *et al*. (2000) The protein data bank. Nucleic Acids Res 28: 235-242.

225. Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, *et al*. (2004) The distribution and query systems of the RCSB protein data bank. Nucleic Acids Res 32: D223-5. 10.1093/nar/gkh096.

226. Li YY, An J, Jones SJ. (2006) A large-scale computational approach to drug repositioning. Genome Inform 17: 239-247.

227. Ogiso H, Ishitani R, Nureki O, Fukai S, Yamanaka M, *et al*. (2002) Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. Cell 110: 775-787.

228. Garrett TP, McKern NM, Lou M, Elleman TC, Adams TE, *et al*. (2002) Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor alpha. Cell 110: 763-773.

229. Ferguson KM, Berger MB, Mendrola JM, Cho HS, Leahy DJ, *et al*. (2003) EGF activates its receptor by removing interactions that autoinhibit ectodomain dimerization. Mol Cell 11: 507-517.

230. Li S, Schmitz KR, Jeffrey PD, Wiltzius JJ, Kussie P, *et al*. (2005) Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. Cancer Cell 7: 301-311. 10.1016/j.ccr.2005.03.003.

231. Stamos J, Sliwkowski MX, Eigenbrot C. (2002) Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. J Biol Chem 277: 46265-46272. 10.1074/jbc.M207135200.

232. Wood ER, Truesdale AT, McDonald OB, Yuan D, Hassell A, *et al*. (2004) A unique structure for epidermal growth factor receptor bound to GW572016 (lapatinib): Relationships among protein conformation, inhibitor off-rate, and receptor activity in

tumor cells. Cancer Res 64: 6652-6659. 10.1158/0008-5472.CAN-04-1168.

233. Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J. (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. Cell 125: 1137-1149. 10.1016/j.cell.2006.05.013.

234. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, *et al*. (2007) Structures of lung cancer-derived EGFR mutants and inhibitor complexes: Mechanism of activation and insights into differential inhibitor sensitivity. Cancer Cell 11: 217-227. 10.1016/j.ccr.2006.12.017.

235. Blair JA, Rauh D, Kung C, Yun CH, Fan QW, *et al*. (2007) Structure-guided development of affinity probes for tyrosine kinases using chemical genetics. Nat Chem Biol 3: 229-238. 10.1038/nchembio866.

236. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, *et al*. (2000) Molecular portraits of human breast tumours. Nature 406: 747-752. 10.1038/35021093.

237. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, *et al*. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98: 10869-10874. 10.1073/pnas.191367098.

238. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, *et al*. (2008) Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clin Cancer Res 14: 1368-1376. 10.1158/1078-0432.CCR-07-1658.

239. Fisher B, Costantino J, Redmond C, Poisson R, Bowman D, *et al*. (1989) A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N Engl J Med 320: 479-484. 10.1056/NEJM198902233200802.

240. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, *et al*. (2005) Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med 353: 1659-1672. 10.1056/NEJMoa052306.

241. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, *et al*. (2007) Triple-negative breast cancer: Clinical features and patterns of recurrence. Clin Cancer Res 13: 4429-4434. 10.1158/1078-0432.CCR-06-3045.

242. Foulkes WD, Smith IE, Reis-Filho JS. (2010) Triple-negative breast cancer. N Engl J Med 363: 1938-1948. 10.1056/NEJMra1001389.

243. Brenton JD, Carey LA, Ahmed AA, Caldas C. (2005) Molecular classification and molecular forecasting of breast cancer: Ready for clinical application? J Clin Oncol 23: 7350-7360. 10.1200/JCO.2005.03.3845.

244. Carriere A, Ray H, Blenis J, Roux PP. (2008) The RSK factors of activating the Ras/MAPK signaling cascade. Front Biosci 13: 4258-4275.

245. Stratford AL, Fry CJ, Desilets C, Davies AH, Cho YY, *et al*. (2008) Y-box binding protein-1 serine 102 is a downstream target of p90 ribosomal S6 kinase in basal-like breast cancer cells. Breast Cancer Res 10: R99. 10.1186/bcr2202.

246. Wu J, Lee C, Yokom D, Jiang H, Cheang MC, *et al*. (2006) Disruption of the Y-box binding protein-1 results in suppression of the epidermal growth factor receptor and HER-2. Cancer Res 66: 4872-4879. 10.1158/0008-5472.CAN-05-3561.

247. Smith JA, Poteet-Smith CE, Xu Y, Errington TM, Hecht SM, *et al*. (2005) Identification of the first specific inhibitor of p90 ribosomal S6 kinase (RSK) reveals an unexpected role for RSK in cancer cell proliferation. Cancer Res 65: 1027-1034.

248. Fisher TL, Blenis J. (1996) Evidence for two catalytically active kinase domains in pp90rsk. Mol Cell Biol 16: 1212-1219.

249. Cardozo T, Totrov M, Abagyan R. (1995) Homology modeling by the ICM method. Proteins 23: 403-414. 10.1002/prot.340230314.

250. Cohen MS, Zhang C, Shokat KM, Taunton J. (2005) Structural bioinformatics-based design of selective, irreversible kinase inhibitors. Science 308: 1318-1321. 10.1126/science1108367.

251. Nguyen TL, Gussio R, Smith JA, Lannigan DA, Hecht SM, *et al*. (2006) Homology model of RSK2 N-terminal kinase domain, structure-based identification of novel RSK2 inhibitors, and preliminary common pharmacophore. Bioorg Med Chem 14: 6097-6105. 10.1016/j.bmc.2006.05.001.

252. Sapkota GP, Cummings L, Newell FS, Armstrong C, Bain J, *et al*. (2007) BI-D1870 is a specific inhibitor of the p90 RSK (ribosomal S6 kinase) isoforms in vitro and in vivo. Biochem J 401: 29-38. 10.1042/BJ20061088.

253. Wade CB, Dorsa DM. (2003) Estrogen activation of cyclic adenosine 5'-monophosphate response element-mediated transcription requires the extracellularly regulated kinase/mitogen-activated protein kinase pathway. Endocrinology 144: 832-838.

254. Koh PO. (2007) Estradiol prevents the injury-induced decrease of 90 ribosomal S6 kinase (p90RSK) and bad phosphorylation. Neurosci Lett 412: 68-72. 10.1016/j.neulet.2006.10.060.

255. Albert JM, Kim KW, Cao C, Lu B. (2006) Targeting the Akt/mammalian target of rapamycin pathway for radiosensitization of breast cancer. Mol Cancer Ther 5: 1183-1189. 10.1158/1535-7163.MCT-05-0400.

256. Lorusso PM, Adjei AA, Varterasian M, Gadgeel S, Reid J, *et al*. (2005) Phase I and pharmacodynamic study of the oral MEK inhibitor CI-1040 in patients with advanced malignancies. J Clin Oncol 23: 5281-5293. 10.1200/JCO.2005.14.415.

257. Shimoi K, Okada H, Furugori M, Goda T, Takase S, *et al*. (1998) Intestinal absorption of luteolin and luteolin 7-O-beta-glucoside in rats and humans. FEBS Lett 438: 220-224.

258. Peng Y, Li C, Chen L, Sebti S, Chen J. (2003) Rescue of mutant *P53* transcription function by ellipticine. Oncogene 22: 4478-4487. 10.1038/sj.onc.1206777.

259. Xu GW, Mawji IA, Macrae CJ, Koch CA, Datti A, *et al*. (2008) A high-content chemical screen identifies ellipticine as a modulator of *P53* nuclear localization. Apoptosis 13: 413-422. 10.1007/s10495-007-0175-4.

260. Arkin MR, Wells JA. (2004) Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. Nat Rev Drug Discov 3: 301-317. 10.1038/nrd1343.

261. Huang SY, Zou X. (2007) Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. Proteins 66: 399-421. 10.1002/prot.21214.

262. Ferrara P, Jacoby E. (2007) Evaluation of the utility of homology models in high throughput docking. J Mol Model 13: 897-905. 10.1007/s00894-007-0207-6.

263. Charifson PS, Corkery JJ, Murcko MA, Walters WP. (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 42: 5100-5109.

264. Maiorov V, Abagyan R. (1998) Energy strain in three-dimensional protein structures. Fold Des 3: 259-269.

265. Yang J, Cron P, Good VM, Thompson V, Hemmings BA, *et al*. (2002) Crystal structure of an activated Akt/protein kinase B ternary complex with GSK3-peptide and AMP-PNP. Nat Struct Biol 9: 940-944. 10.1038/nsb870.

266. Law JH, Li Y, To K, Wang M, Astanehe A, *et al*. (2010) Molecular decoy to the Y-box binding protein-1 suppresses the growth of breast and prostate cancer cells whilst sparing normal cell viability. PLoS One 5: e12661. 10.1371/journal.pone.0012661.

267. Sapkota GP, Cummings L, Newell FS, Armstrong C, Bain J, *et al*. (2007) BI-D1870 is a specific inhibitor of the p90 RSK (ribosomal S6 kinase) isoforms in vitro and in vivo. Biochem J 401: 29-38. 10.1042/BJ20061088.

268. Wu J, Lee C, Yokom D, Jiang H, Cheang MC, *et al*. (2006) Disruption of the Y-box binding protein-1 results in suppression of the epidermal growth factor receptor and

HER-2. Cancer Res 66: 4872-4879. 10.1158/0008-5472.CAN-05-3561.

269. Law JH, Habibi G, Hu K, Masoudi H, Wang MY, *et al*. (2008) Phosphorylated insulin-like growth factor-i/insulin receptor is present in all breast cancer subtypes and is related to poor survival. Cancer Res 68: 10238-10246. 10.1158/0008-5472.CAN-08-2755.

270. Sutherland BW, Kucab J, Wu J, Lee C, Cheang MC, *et al*. (2005) Akt phosphorylates the Y-box binding protein 1 at Ser102 located in the cold shock domain and affects the anchorage-independent growth of breast cancer cells. Oncogene 24: 4281-4292. 10.1038/sj.onc.1208590.

271. Ikuta M, Kornienko M, Byrne N, Reid JC, Mizuarai S, *et al*. (2007) Crystal structures of the N-terminal kinase domain of human RSK1 bound to three different ligands: Implications for the design of RSK1 specific inhibitors. Protein Sci 16: 2626-2635. 10.1110/ps.073123707.

272. Malakhova M, Kurinov I, Liu K, Zheng D, D'Angelo I, *et al*. (2009) Structural diversity of the active N-terminal kinase domain of p90 ribosomal S6 kinase 2. PLoS One 4: e8044. 10.1371/journal.pone.0008044.

273. Malakhova M, Tereshko V, Lee SY, Yao K, Cho YY, *et al*. (2008) Structural basis for activation of the autoinhibitory C-terminal kinase domain of p90 RSK2. Nat Struct Mol Biol 15: 112-113. 10.1038/nsmb1347.

274. Bain J, Plater L, Elliott M, Shpiro N, Hastie CJ, *et al*. (2007) The selectivity of protein kinase inhibitors: A further update. Biochem J 408: 297-315. 10.1042/BJ20070797.

275. Traxler P, Allegrini PR, Brandt R, Brueggen J, Cozens R, *et al*. (2004) AEE788: A dual family epidermal growth factor receptor/ErbB2 and vascular endothelial growth factor receptor tyrosine kinase inhibitor with antitumor and antiangiogenic activity. Cancer Res 64: 4931-4941. 10.1158/0008-5472.CAN-03-3681.

276. Jemal A, Bray F, Center MM, Ferlay J, Ward E, *et al*. (2011) Global cancer statistics. CA Cancer J Clin 61: 69-90. 10.3322/caac.20107.

277. Kamb A. (2005) What's wrong with my cancer models? Nat Rev Drug Discov 4: 161-165. 10.1038/nrd1635.

278. Kamb A, Wee S, Lengauer C. (2007) Why is cancer drug discovery so difficult? Nat Rev Drug Discov 6: 115-120. 10.1038/nrd2155.

279. Fagerlund TH, Braaten O. (2001) No pain relief from codeine...? an introduction to pharmacogenomics. Acta Anaesthesiol Scand 45: 140-149.

280. Gazdar AF. (2009) Personalized medicine and inhibition of EGFR signaling in lung cancer. N Engl J Med 361: 1018-1020. 10.1056/NEJMe0905763.

281. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, *et al*. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98: 10869-10874. 10.1073/pnas.191367098.

282. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, *et al*. (2005) Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res 11: 5678-5685. 10.1158/1078-0432.CCR-04-2421.

283. Linehan WM, Pinto PA, Srinivasan R, Merino M, Choyke P, *et al*. (2007) Identification of the genes for kidney cancer: Opportunity for disease-specific targeted therapeutics. Clin Cancer Res 13: 671s-679s. 10.1158/1078-0432.CCR-06-1870.

284. Kurman RJ, Shih I. (2010) The origin and pathogenesis of epithelial ovarian cancer: A proposed unifying theory. Am J Surg Pathol 34: 433-443. 10.1097/PAS.0b013e3181cf3d79.

285. Sotiriou C, Pusztai L. (2009) Gene-expression signatures in breast cancer. N Engl J Med 360: 790-800. 10.1056/NEJMra0801289.

286. Bubendorf L, Grilli B, Sauter G, Mihatsch MJ, Gasser TC, *et al*. (2001) Multiprobe FISH for enhanced detection of bladder cancer in voided urine specimens and bladder washings. Am J Clin Pathol 116: 79-86. 10.1309/K5P2-4Y8B-7L5A-FAA9.

287. Krizman DB, Wagner L, Lash A, Strausberg RL, Emmert-Buck MR. (1999) The cancer genome anatomy project: EST sequencing and the genetics of cancer progression. Neoplasia 1: 101-106.

288. Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, *et al*. (1999) A public database for gene expression in human cancers. Cancer Res 59: 5403-5407.

289. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. (1995) Serial analysis of gene expression. Science 270: 484-487.

290. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, *et al*. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153-158. 10.1038/nature05610.

291. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, *et al*. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455: 1069-1075. 10.1038/nature07423.

292. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455:

1061-1068. 10.1038/nature07385.

293. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, *et al*. (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314: 268-274. 10.1126/science.1133427.

294. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, *et al*. (2007) The genomic landscapes of human breast and colorectal cancers. Science 318: 1108-1113. 10.1126/science.1145720.

295. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, *et al*. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321: 1801-1806. 10.1126/science.1164368.

296. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59. 10.1038/nature07517.

297. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, *et al*. (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7: 246. 10.1186/1471-2164-7-246.

298. Hashimoto S, Qu W, Ahsan B, Ogoshi K, Sasaki A, *et al*. (2009) High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. PLoS One 4: e4108. 10.1371/journal.pone.0004108.

299. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, *et al*. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques 45: 81-94. 10.2144/000112900.

300. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, *et al*. (2009) Mutational evolution in a lobular breast tummy profiled at single nucleotide resolution. Nature 461: 809-813. 10.1038/nature08489.

301. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, *et al*. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66-72. 10.1038/nature07485.

302. de Diego JI, Bernaldez R, Prim MP, Hardisson D. (1996) Polymorphous low-grade adenocarcinoma of the tongue. J Laryngol Otol 110: 700-703.

303. Kennedy KS, Healy KM, Taylor RE, Strom CG. (1987) Polymorphous low-grade adenocarcinoma of the tongue. Laryngoscope 97: 533-536.

304. Unal M, Polat A, Akbas Y, Pata Y. (2004) Polymorphous low-grade adenocarcinoma of the tongue. Auris Nasus Larynx 31: 85-88. 10.1016/j.anl.2003.08.001.

305. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5: 621-628. 10.1038/nmeth.1226.

306. Rutherford S, Hampton GM, Frierson HF, Moskaluk CA. (2005) Mapping of candidate tumor suppressor genes on chromosome 12 in adenoid cystic carcinoma. Lab Invest 85: 1076-1085. 10.1038/labinvest.3700314.

307. Gillenwater A, Hurr K, Wolf P, Batsakis JG, Goepfert H, *et al*. (1997) Microsatellite alterations at chromosome 8q loci in pleomorphic adenoma. Otolaryngol Head Neck Surg 117: 448-452.

308. Sahlin P, Mark J, Stenman G. (1994) Submicroscopic deletions of 3p sequences in pleomorphic adenomas with t(3;8)(p21;q12). Genes Chromosomes Cancer 10: 256-261.

309. Stenman G, Sandros J, Mark J, Edstrom S. (1989) Partial 6q deletion in a human salivary gland adenocarcinoma. Cancer Genet Cytogenet 39: 153-156.

310. Kanehisa M, Goto S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.

311. Zbuk KM, Eng C. (2007) Cancer phenomics: RET and PTEN as illustrative models. Nat Rev Cancer 7: 35-45. 10.1038/nrc2037.

312. Lanzi C, Cassinelli G, Nicolini V, Zunino F. (2009) Targeting RET for thyroid cancer therapy. Biochem Pharmacol 77: 297-309. 10.1016/j.bcp.2008.10.033.

313. Woo J, Lee J, Kim MS, Jang SJ, Sidransky D, *et al*. (2008) The effect of aquaporin 5 overexpression on the ras signaling pathway. Biochem Biophys Res Commun 367: 291-298. 10.1016/j.bbrc.2007.12.073.

314. Hennessy BT, Smith DL, Ram PT, Lu Y, Mills GB. (2005) Exploiting the PI3K/AKT pathway for cancer drug discovery. Nat Rev Drug Discov 4: 988-1004. 10.1038/nrd1902.

315. Lee JI, Soria JC, Hassan KA, El-Naggar AK, Tang X, *et al*. (2001) Loss of PTEN expression as a prognostic marker for tongue cancer. Arch Otolaryngol Head Neck Surg 127: 1441-1445.

316. Gu J, Tamura M, Yamada KM. (1998) Tumor suppressor PTEN inhibits integrin- and growth factor-mediated mitogen-activated protein (MAP) kinase signaling pathways. J Cell Biol 143: 1375-1383.

317. She QB, Solit D, Basso A, Moasser MM. (2003) Resistance to gefitinib in PTEN-null HER-overexpressing tumor cells can be overcome through restoration of PTEN function or pharmacologic modulation of constitutive phosphatidylinositol 3'-kinase/Akt pathway signaling. Clin Cancer Res 9: 4340-4346.

318. Yamasaki F, Johansen MJ, Zhang D, Krishnamurthy S, Felix E, *et al*. (2007) Acquired resistance to erlotinib in A-431 epidermoid cancer cells requires down-regulation of MMAC1/PTEN and up-regulation of phosphorylated akt. Cancer Res 67: 5779-5788. 10.1158/0008-5472.CAN-06-3020.

319. Albitar L, Carter MB, Davies S, Leslie KK. (2007) Consequences of the loss of *P53*, RB1, and PTEN: Relationship to gefitinib resistance in endometrial cancer. Gynecol Oncol 106: 94-104. 10.1016/j.ygyno.2007.03.006.

320. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, *et al*. (2000) New guidelines to evaluate the response to treatment in solid tumors. european organization for research and treatment of cancer, national cancer institute of the united states, national cancer institute of canada. J Natl Cancer Inst 92: 205-216.

321. John A, Tuszynski G. (2001) The role of matrix metalloproteinases in tumor angiogenesis and tumor metastasis. Pathol Oncol Res 7: 14-23.

322. Patarroyo M, Tryggvason K, Virtanen I. (2002) Laminin isoforms in tumor invasion, angiogenesis and metastasis. Semin Cancer Biol 12: 197-207. 10.1016/S1044-579X(02)00023-8.

323. Alexander W. (2011) Inhibiting the akt pathway in cancer treatment: Three leading candidates. P T 36: 225-227.

324. Houghton PJ. (2010) Everolimus. Clin Cancer Res 16: 1368-1372. 10.1158/1078-0432.CCR-09-1314.

325. Duncia JV, Santella JB,3rd, Higley CA, Pitts WJ, Wityak J, *et al*. (1998) MEK inhibitors: The chemistry and biological activity of U0126, its analogs, and cyclization products. Bioorg Med Chem Lett 8: 2839-2844.

326. Adjei AA, Cohen RB, Franklin W, Morris C, Wilson D, *et al*. (2008) Phase I pharmacokinetic and pharmacodynamic study of the oral, small-molecule mitogen-activated protein kinase kinase 1/2 inhibitor AZD6244 (ARRY-142886) in patients with advanced cancers. J Clin Oncol 26: 2139-2146. 10.1200/JCO.2007.14.4956.

327. Huang D, Ding Y, Zhou M, Rini BI, Petillo D, *et al*. (2010) Interleukin-8 mediates resistance to antiangiogenic agent sunitinib in renal cell carcinoma. Cancer Res 70: 1063-1071. 10.1158/0008-5472.CAN-09-3965.

328. Luppi F, Longo AM, de Boer WI, Rabe KF, Hiemstra PS. (2007) Interleukin-8 stimulates cell proliferation in non-small cell lung cancer through epidermal growth factor receptor transactivation. Lung Cancer 56: 25-33. 10.1016/j.lungcan.2006.11.014.

329. Dudek AZ, Zolnierek J, Dham A, Lindgren BR, Szczylik C. (2009) Sequential therapy with sorafenib and sunitinib in renal cell carcinoma. Cancer 115: 61-67. 10.1002/cncr.24009.

330. Herrmann E, Marschner N, Grimm MO, Ohlmann CH, Hutzschenreuter U, *et al*. (2011) Sequential therapies with sorafenib and sunitinib in advanced or metastatic renal cell carcinoma. World J Urol 29: 361-366. 10.1007/s00345-011-0673-4.

331. Sablin MP, Negrier S, Ravaud A, Oudard S, Balleyguier C, *et al*. (2009) Sequential sorafenib and sunitinib for renal cell carcinoma. J Urol 182: 29-34; discussion 34. 10.1016/j.juro.2009.02.119.

332. Buchler T, Klapka R, Melichar B, Brabec P, Dusek L, *et al*. (2011) Sunitinib followed by sorafenib or vice versa for metastatic renal cell carcinoma--data from the czech registry. Ann Oncol . 10.1093/annonc/mdr065.

333. Tanaka T, Watanabe T, Kazama Y, Tanaka J, Kanazawa T, *et al*. (2006) Chromosome 18q deletion and Smad4 protein inactivation correlate with liver metastasis: A study matched for T- and N- classification. Br J Cancer 95: 1562-1567. 10.1038/sj.bjc.6603460.

334. Rini BI, Wilding G, Hudes G, Stadler WM, Kim S, *et al*. (2009) Phase II study of axitinib in sorafenib-refractory metastatic renal cell carcinoma. J Clin Oncol 27: 4462-4468. 10.1200/JCO.2008.21.7034.

335. Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, *et al*. (2010) Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat Genet 42: 181-185. 10.1038/ng.518.

336. Shah SP, Kobel M, Senz J, Morin RD, Clarke BA, *et al*. (2009) Mutation of FOXL2 in granulosa-cell tumors of the ovary. N Engl J Med 360: 2719-2729. 10.1056/NEJMoa0902542.

337. Li H, Ruan J, Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858. 10.1101/gr.078212.108.

338. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, *et al*. (2001) dbSNP: The NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

339. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-876.

10.1038/nature06884.

340. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, *et al*. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254. 10.1371/journal.pbio.0050254.

341. Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, *et al*. (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22: e431-9. 10.1093/bioinformatics/btl238.

342. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, *et al*. (2009) Ensembl 2009. Nucleic Acids Res 37: D690-7. 10.1093/nar/gkn828.

343. Benjamini Y, Hochberg Y. (1995) Controlling the false positive discovery rate: A practical and powerful approach to multiple testing. Royal Stat Soc 57: 289-300.

344. Hulsen T, de Vlieg J, Alkema W. (2008) BioVenn - a web application for the comparison and visualization of biological lists using area-proportional venn diagrams. BMC Genomics 9: 488. 10.1186/1471-2164-9-488.

345. Terry J, Saito T, Subramanian S, Ruttan C, Antonescu CR, *et al*. (2007) TLE1 as a diagnostic immunohistochemical marker for synovial sarcoma emerging from gene expression profiling studies. Am J Surg Pathol 31: 240-246. 10.1097/01.pas.0000213330.71745.39.

346. Terry J, Barry TS, Horsman DE, Hsu FD, Gown AM, *et al*. (2005) Fluorescence in situ hybridization for the detection of t(X;18)(p11.2;q11.2) in a synovial sarcoma tissue microarray using a breakapart-style probe. Diagn Mol Pathol 14: 77-82.

347. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, *et al*. (2009) Circos: An information aesthetic for comparative genomics. Genome Res 19: 1639-1645. 10.1101/gr.092759.109.

348. Oshiro C, Bradley EK, Eksterowicz J, Evensen E, Lamb ML, *et al*. (2004) Performance of 3D-database molecular docking studies into homology models. J Med Chem 47: 764-767. 10.1021/jm0300781.

349. Venclovas C, Ginalski K, Kang C. (2004) Sequence-structure mapping errors in the PDB: OB-fold domains. Protein Sci 13: 1594-1602. 10.1110/ps.04634604.

350. Ament PW, Bertolino JG, Liszewski JL. (2000) Clinically significant drug interactions. Am Fam Physician 61: 1745-1754.

351. Roskoski R,Jr. (2007) Sunitinib: A VEGF and PDGF receptor protein kinase and angiogenesis inhibitor. Biochem Biophys Res Commun 356: 323-328. 10.1016/j.bbrc.2007.02.156.

352. Plaza-Menacho I, Mologni L, Sala E, Gambacorti-Passerini C, Magee AI, *et al*. (2007) Sorafenib functions to potently suppress RET tyrosine kinase activity by direct enzymatic inhibition and promoting RET lysosomal degradation independent of proteasomal targeting. J Biol Chem 282: 29230-29240. 10.1074/jbc.M703461200.

353. Wilhelm SM, Carter C, Tang L, Wilkie D, McNabola A, *et al*. (2004) BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis. Cancer Res 64: 7099-7109. 10.1158/0008-5472.CAN-04-1443.

354. To K, Fotovati A, Reipas KM, *et al*. (2010) Y-box binding protein-1 induces the expression of CD44 and CD49f leading to enhanced self-renewal, mammosphere growth, and drug resistance. Cancer Res 70: 2840-51.24

355. Schapira M, Abagyan R, Totrov M. (2002) Structural model of nicotinic acetylcholine receptor isotypes bound to acetylcholine and nicotine. BMC Struct Biol 2:1. 10.1186/1472-6807-2-1

356. Marsden B, Abagyan A. SAD—a normalized structural alignment database: improving sequence–structure alignments. Bioinformatics. 20(15): 2333-2344. 10.1093/bioinformatics/bth244.

357. Cardozo T, Totrov M, Abagyan A. Homology modeling by the ICM method. Proteins. 23(3): 403-414. 10.1002/prot.340230314

Appendices

Appendix A Top homology modeling templates for RSK1.

Top template options for RSK homology modeling as determined by the SWISS-MODEL server. These PDB structures are selected by sequence identity to the RSK kinase domain. 14 structures of AKT are omitted from this table; they either 1) have similar sequence identities and resolutions 2) are complexed with a small molecule inhibitor only, or 3) were deposited in PDB in 2009-2011 which was after the initial modeling.

PDB ID	Resolution (Å)	Protein name	Sequence identity
3a62	2.4	p70S6K1	55%
3a60	2.8	p70S6K1	55%
1vzo	1.8	MSK	54%
3a61	3.4	p70S6K1	53%
3e87	2.3	AKT2	45%
3d0e	2.0	AKT2	45%
3e88	2.5	AKT2	45%
1o6k	1.7	AKT1 (S474D)	45%
2jdo	1.8	AKT2	45%
3e8d	2.7	AKT2	45%
1o6l	1.6	AKT1 (PIF)	45%
3hdm	2.6	SGK1	45%
3iw4	2.8	PKCalpha	43%
3dne	2.0	PKA	38%

Appendix B Comparison of sunitinib and sorafenib polypharmacology

a) Comparison of the sorafenib and sunitinib targets based on 317 direct kinase binding assays from [51] using the KinomeScan technology (Ambit Biosciences, Massachusetts, USA). Only IC_{50} values below 100nM are shown. Shared targets are shaded in grey at the bottom of the table. Surprisingly, sunitinib inhibited 43 of 317 kinases with IC_{50}'s below 100nM and 200 of 317 kinases with IC_{50}'s within 10μM.

Target	**sorafenib IC_{50} (nM)**	**sunitinib IC_{50} (nM)**
AAK1		11
AMPK-alpha1		19
AMPK-alpha2		89
ARK5		48
AXL		9
BLK		65
CAMK2A		80
CLK1		22
CLK2		20
CLK4		29
CSNK1D		15
CSNK1E		13
CSNK2A1		81
DAPK3		22
DRAK1		1
GAK		20
ITK		13
JAK3		49
LKB		38
LOK		19
MAP4K1		16
MAP4K5		41
MERTK		25
MLCK		23
STK3		56
STK4		19
STK24		63
MYLK		49
PAK3		16
PHKG1		5.5
PHKG2		5.9
PIP5K2B		39
PTK2B		82
RIOK1		35
RPS6KA2		17
RPS6KA4		96

Target	sorafenib IC_{50} (nM)	sunitinib IC_{50} (nM)
RPS6KA5		28
SGK085		15
STK33		17
SLK		56
TNIK		25
TTK		63
TYRO3		49
Raf-1	230	
BRAF	540	
TIE1	68	
DDR1	1.5	
DDR2	6.6	
MAPK15	46	
ZAK	6.3	
CSF1R	28	2
FLT1	31	1.8
FLT3	13	0.47
KIT	31	0.37
PDGFRA	62	0.79
PDGFRB	37	0.075
RET	13	12
VEGFR2	59	1.5

b) Kinase targets widely established as major sunitinib targets. [351]

Target	**sunitinib IC_{50} (nM)**
VEGFR1	15
VEGFR2	38
VEGR3	30
PDGFRA	69
PDGFRB	55
CSF1R	35
FLT3	21
KIT	1
RET	224
FGFR1	675

c) Kinase targets widely established as major sorafenib targets. [352, 353]

Target	**sorafenib IC_{50} (nM)**
RET	5.9
RAF-1	6
BRAF	22
VEGFR2	90
murine VEGFR3	20
murine PDGFRB	57
FLT3	58
KIT	68
FGFR1	580

Appendix C RNA-Seq libraries included in the compendium.

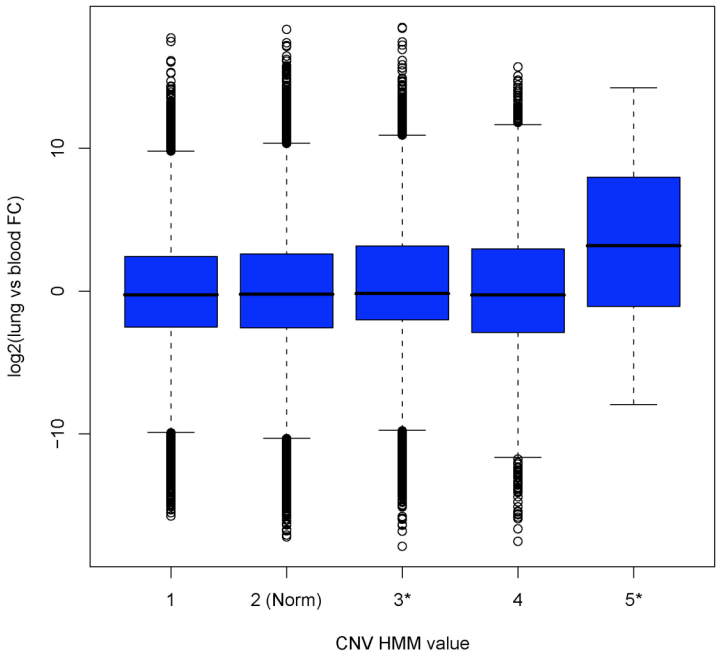
The compendium is comprised of 50 RNA-Seq libraries including 19 cell lines and 31 primary samples representing at least 19 different tissues and 25 tumor types as well as 6 normal or benign samples. Cell line names are listed in brackets under 'Tissue' where applicable. Otherwise, all libraries were derived from primary tumors.

Tissue (Cell Line: If applicable)	**Description**	**Gender**
Bone marrow	Acute Lymphoblastic Leukemia	Unknown
Bone marrow	Acute Lymphoblastic Leukemia	Unknown
Brain	Oligodendroglioma	Unknown
Brain	Oligodendroglioma	Unknown
Brain	Oligodendroglioma	Unknown
Brain (NB88)	Neuroblastoma	Unknown
Brain/Bone Marrow (NB122L)	Neuroblastoma, stage 4, bone marrow metastases	Male
Brain/Bone Marrow (NB153)	Neuroblastoma, stage 4, bone marrow metastases	Unknown
Breast	Breast Tumor	Female
Breast	Breast Tumor	Female
Breast	Breast Tumor	Female
Breast (BT474-M1)	Solid, invasive ductal carcinoma	Female
Breast (HS-578T)	Aneuploid epithelial breast carcinoma	Female
Breast (SUM149)	Breast carcinoma	Female
Colon (HCT116)	Colon carcinoma	Male
Colon (MIP101)	Colon carcinoma	Male
Embryonic stem cells	Normal, undifferentiated	Male
Foreskin (FS210)	Normal	Male
Foreskin (FS248)	Normal	Male
Foreskin (FS253)	Normal skin-derived precursor cells	Male
Gastrointestinal Tract	Lymphoma	Female
Lung	Lung tumor	Female
Lung	Lung tumor	Female
Lung (PC9)	Lung adenocarcinoma	Unknown
Lymph nodes	Lymphoma	Male
Lymph nodes	Lymphoma	Male
Lymph nodes	Primary mediastinal B cell lymphoma	Female
Mononuclear blood cells	Acute Lymphoblastic Leukemia	Male
Mononuclear peripheral blood cells	Acute Lymphoblastic Leukemia	Male

Tissue (Cell Line: If applicable)	**Description**	**Gender**
Ovary	Endometroid ovarian cancer	Female
Ovary	High grade clear cell ovarian tumor	Female
Ovary	High grade serous cancer	Female
Ovary	Mucinous ovarian cancer	Female
Ovary	Small cell hypercalemic ovarian cancer	Female
Ovary	Granulosa cell ovarian tumor	Female
Ovary (BIN67)	Ovarian small cell carcinoma	Female
Ovary (SBOT 3.1)	Low grade serous ovarian tumor	Female
Pancreas (CAPAN-1)	Pancreatic adenocarcinoma	Male
Pelvis, Right	Epithelioid sarcoma	Unknown
Peripheral blood	Acute Lymphoblastic Leukemia	Unknown
Peripheral blood	Acute Lymphoblastic Leukemia	Unknown
Peritoneal effusion (SU-DHL-6)	B-cell Non-Hodgkin Lymphoma	Male
Pleural effusion (Karpas 1106P)	Primary Mediastinal B cell lymphoma	Female
Pleural effusion (KM-H2)	Hodgkin lymphoma, mixed cellulariity	Male
Skin (A431)	Epidermoid carcinoma	Female
Spleen	Lymphoma	Female
Thigh	Sarcoma	Unknown
Tonsil	Benign (CD77+ normoblasts)	Unknown
Tonsil	Benign (centroblast cells)	Unknown
Tonsil	Lymphoma	Male

Appendix D Copy number varation (CNV) compared to differential expression.

a) Boxplot of CNV versus differential expression value for lung metastasis relative to blood.
* Indicates significant p-value for comparison to normal (CNV = 2).



b) Boxplot of CNV versus differential expression value for skin metastasis relative to blood.
* Indicates significant p-value for comparison to normal (CNV = 2).