

BIOINFORMATICS OF TARGETED THERAPEUTICS AND
APPLICATIONS IN DRUG DISCOVERY

Qin Chu

(B.Sc. (Hons.), NUS)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE
SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Qin Chu

18 August 2014

Acknowledgements

First and foremost, I would like to express my heartfelt appreciation to my supervisor, Professor Chen Yu Zong, for his invaluable guidance on my research projects and inspiring encouragement throughout the years.

Many thanks to my thesis advisory committee members A/Prof Chandra Shekhar Verma and Dr Yap Chun Wei for providing insightful comments on my research and devoting their time to be my qualification examination examiners.

I wish to thank all the previous and current members of the BIDD group for the valuable discussions and timely help in the past four years. And my sincere appreciation goes to my friends for their encouragement and trust all the time.

Lastly but most importantly, my deepest appreciation and love is dedicated to my family. Your unconditional love is my source of courage and happiness.

Table of Contents

Declaration.....	1
Acknowledgements.....	i
Summary	vii
List of Tables.....	xi
List of Figures	xiv
List of Abbreviations.....	xvii
List of Publications	xix
Chapter 1: Introduction	1
1.1 Overview of targeted therapeutics in modern drug discovery	1
1.2 The importance of multi-target therapeutics	6
1.3 More personalized targeted therapeutics driven by biomarkers	7
1.4 Bioinformatics methods for analysis of targeted therapeutics	10
1.4.1 The update of therapeutic target database to serve as an integrated information platform of targeted therapeutics	10
1.4.2 Machine learning methods to predict multi-target agents from large chemical libraries	12
1.4.3 Clustering method to analyse the distribution patterns of targeted drugs in target-specific chemical space	15
1.4.4 Systematic analysis to study synergistic combinations of natural	

products as potential sources of multi-targeted therapeutics	18
1.4.5 Analysis of biomarker for personalized medicine	20
1.5: Outline of thesis	22
Chapter 2 Update of therapeutic target database as an integrated source of targeted therapeutics data.....	25
2.1 Statistics of updated targeted therapeutics in TTD	26
2.2 Materials and methods.	29
2.2.1 Data collection method	29
2.2.2. Data sources	31
2.3. Data in TTD and ways to access them.....	34
2.3.1 Overall search and download options	34
2.3.2 Targets and drugs	40
2.3.3 Biomarkers	46
2.3.4 Multi-target agents and drug combinations	51
2.3.5 International Classification of Disease	52
2.4 Future work.....	57
Chapter 3: Methods to learn from known drugs and inhibitors for the design of multi-target small molecule drugs.....	59
3.1 Evaluation of Hit and Target Selection Performance of Machine Learning Multi-Target Virtual Screening Methods	59
3.1.1 Method	60

3.1.2 Results and discussion	71
3.1.3 Future work.....	81
3.2 Hints of drug prolific regions and properties by clustering drugs in the target-specific chemical space	83
3.2.1 Data collection and method	84
3.2.2 Preliminary results	87
Chapter 4. Specific multi-target modes identified by analysing synergistic natural product combination	98
4.1 Method	101
4.2 Results and discussion	103
4.2.1 Comparison of the potencies of natural products and drugs in cell-based assays	103
4.2.3 Potency enhancing molecular modes of natural product combinations	107
4.3 Summary	117
Chapter 5: Personalized targeted therapeutics driven by biomarkers	120
5.1 More refined classification of patient subpopulations for personalized targeted therapeutics	121
5.2 Non-invasive biomarker and their applications to healthcare.....	126
5.2.1 Background	126
5.2.2 Evaluation of new biomarker-detection technologies.....	130
5.2.3 The relevance and accuracies of the non-invasive molecular	

biomarkers for mhealth applications.....	131
5.2.4 A digitally-coded biomarker, disease and therapeutic information processing system	133
5.2.5 Future work.....	134
Chapter 6: Concluding remarks	154
Bibliography	159
Appendices.....	171

Summary

The modern rational drug discovery process starts with the hypothesis that modulation of certain targets may exert therapeutic value and therapeutics directed at those targets are then developed to combat diseases. In this big data era, the large and complex collection of various targeted therapeutics data call for efficient data management and analysis methods. The development of databases to curate, store, integrate and retrieve data and methods to analyze and visualize data are of importance and practical use to increase the success rate of drug discovery.

This work starts with the update of the Therapeutic Target Database (TTD), which serves as a comprehensive, reliable and integrated information source of therapeutics data, including drug targets, drug molecules, natural products and biomarkers. The search tools implemented by the International Classification of Disease (ICD) codes were added to link and retrieve the target, biomarker and drug information. Biomarker information was newly added to the TTD and the data contents were significantly expanded. The updated TTD database enables more convenient data access and will facilitate the discovery, investigation, application, monitoring and management of targeted therapeutics.

An important strategy in targeted therapeutics is the use of multi-target therapeutics such as multi-target drugs and drug combinations, which are more efficacious and

less prone to resistance than single-target drugs for heterogenetic diseases like cancer. To facilitate the multi-target drug discovery, bioinformatics methods such as machine learning methods to predict multi-target inhibitors, clustering method to look for drug prolific regions and properties and systematic analysis of synergistic natural product combinations were developed based on the information from TTD.

Three machine learning methods, support vector machine (SVM), K-Nearest Neighbor(kNN) and probabilistic neural network (PNN) were developed as virtual screening tools to predict dual-target inhibitors from large chemical libraries. Models of 29 targets pairs with varying similarity levels between their drug-binding domains were developed and showed good performance with reasonably high yields and low false hit rates. But the target selectivity performance of these VS tools needs improvement. In search of clues to further modify the virtual hits for drug development, a hierarchical clustering method was proposed to cluster known drugs in the chemical space. Preliminary investigation seemed to hint some drug prolific regions and properties.

Moreover, natural product combinations was systematically analyzed to learn novel multi-target mechanisms. And it was found that most of the evaluated natural products and combinations are sub-potent to drugs. Sub-potent natural products can be assembled into combinations of drug level potency, though at relatively low

probabilities. Distinguished multi-target modes were identified and could shed light to the design of multi-target therapeutics.

In view of the current shift of drug development focus to more personalized targeted therapeutics, the collected comprehensive set of biomarkers and the relevant information were systematically analyzed. The analysis of current biomarkers in TTD with respect to ICD disease classifications suggested that biomarker (especially multi-marker), target and drug information may be incorporated into revised ICD codes for coding disease subclasses and refining patient and drug-response sub-populations for personalized treatment. In addition, the feasibility of utilizing non-invasive biomarkers for mobile health applications was discussed.

List of Tables

Chapter 2

Table 2. 1 Statistics of the drug targets, drugs and their structure and potency data in 2014 version of TTD database.	28
Table 2. 2 List of ICD-9-CM and ICD-10-CM code blocks and the corresponding classes of diseases and related health problems.....	54

Chapter 3

Table 3. 1 Datasets of individual-target and multi-target inhibitors of the target-pairs used for developing and testing machine learning multi-target inhibitor virtual screening tools. Additional sets of 17 million PubChem compounds and 168,000 MDDR active compounds were also used for the test.	62
Table 3. 2 Virtual screening performance of combinatorial SVMs for identifying dual-target inhibitors of high similarity target pairs	71
Table 3. 3 Overall statistics of drugs, inhibitors, structurally similar approved drugs directed to other drugs, similar bioactive ChEMBL compounds and similar non-bioactive Pubchem compounds to be clustered.	84
Table 3. 4 Statistics of drugs, inhibitors, structurally similar approved drugs directed to other drugs, similar bioactive ChEMBL compounds and similar non-bioactive Pubchem compounds in each subtree.	89

Chapter 4

Table 4. 1 The targets and potency-enhancing synergistic molecular modes of the anticancer combination of Tetraarsenic tetrasulfide, Indirubin, and Tanshinone IIA (anticancer synergism reported in literature(161))).	109
Table 4. 2 The targets and potency-enhancing synergistic molecular modes of the anti-rotavirus combination of Theaflavin, Theaflavin-3-monogallate, Theaflavin-3'-monogallate, and Theaflavin-3,3' digallate (anti-rotavirus synergism reported in literature (162))......	112
Table 4. 3 Expression profiles of the primary targets and some of the potency-enhancing secondary targets of the selected natural product combinations in specific patient groups	118

Chapter 5

Table 5. 1 Approved and clinically tested biomarkers for facilitating the prescription of a particular drug to specific patient subpopulation.....	121
Table 5. 2 Examples of diseases and their molecular or cell-based subtypes, ICD codes (marked as NA if unavailable), and the availability (A) or unavailability (NA) of the corresponding diagnostic, prognostic and theragnostic biomarkers and if one or more biomarkers are in clinical use or trial	123
Table 5. 3 New biomarker-detection technologies.	135

Table 5. 4 Diseases covered by non-invasive molecular biomarkers	138
---	-----

Table 5. 5 Conventional test performance	152
---	-----

List of Figures

Chapter 1

Figure 1. 1 Modern drug discovery process.....	2
---	---

Chapter 2

Figure 2. 1 Screenshot of TTD home page.	36
Figure 2. 2 Screenshot of TTD customized search	37
Figure 2. 3 Screenshot of TTD customized search of biomarkers	38
Figure 2. 4 Screenshot of database download page in TTD.	39
Figure 2. 5 Screenshots of detailed information page of ABL1 target.....	43
Figure 2. 6 Screenshots of detailed information page of drug Imatinib.....	46
Figure 2. 7 Screenshot of biomarker detail information of p53.....	50

Chapter 3

Figure 3. 1 Dual model performance of three machine learning methods.....	75
Figure 3. 2 Selectivity of three methods against individual-target inhibitors	78
Figure 3. 3 The virtual hit rates of three machine learning methods to screen MDDR80	
Figure 3. 4 Distribution graph of FLT3 subtree ID 10, labelled according to potency	
values. The labels are colored as follows: red for Approved drug, purple for Phase	
III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for	
inhibitors, grey for similar ChEMBL compounds, pale grey for similar Pubchem	
compounds.	94

Figure 3. 5 Distribution graph of FLT3 subtree ID 10, labelled according to ligand efficiency values. The labels are colored as follows: red for Approved drug, purple for Phase III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for inhibitors, grey for similar ChEMBL compounds, pale grey for similar Pubchem compounds95

Figure 3. 6 Distribution graph of FLT3 subtree ID 10, labelled according to the calculated clogP values. The labels are colored as follows: red for Approved drug, purple for Phase III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for inhibitors, grey for similar ChEMBL compounds, pale grey for similar Pubchem compounds96

Figure 3. 7 Distribution graph of FLT3 subtree ID 10, labelled according to molecular weight. The labels are colored as follows: red for Approved drug, purple for Phase III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for inhibitors, grey for similar ChEMBL compounds, pale grey for similar Pubchem compounds97

Chapter 4

Figure 4. 1 Potency distribution profiles of 88 and 650 anticancer drugs and natural products..... 104

Figure 4. 2 Potency distribution profiles of 102, 609 and 99 antibacterial drugs, natural products (NPs) and NP extracts. 105

Figure 4. 3 Synergism level of 124 synergistic NP combinations. VSS, SS, S, MS,

sS: very strong, strong, normal, moderate, slight synergism, NA: nearly additive, SA,

MA: slight, moderate antagonism..... 106

Figure 4. 4 The potency improvement profile of the constituent NPs..... 107

Chapter 5

Figure 5. 1 Disease-coverage profiles of the biomarkers..... 128

List of Abbreviations

ACE	Angiotensin converting enzyme
ATC	Anatomical Therapeutic Chemical
B2AR	Beta-2 adrenoreceptor
CAS	Chemical Abstracts Service
CDK	Cyclin-dependent kinase
CI	Combination index
COX2	Cyclooxygenase-2
CV	Cross validation
DA1R	Dopamine D1 receptor
DRI	Dose reduction index
ELISA	Enzyme-linked immunosorbent assay
FDA	Food and drug administration
FGFR	Fibroblast growth factor receptor
FLT3	Fms-related tyrosine kinase 3
FN	False negatives
FP	False positives
GI	Growth inhibition
HTS	High-throughput screening
IC	Inhibitory concentration
ICD	International Classification of Disease
iTOL	Interactive tree of life
KEGG	Kyoto Encyclopedia of Genes and Genomes
kNN	k-nearest neighbor
LE	Ligand efficiency
MCC	Matthews Correlation Coefficient
MDDR	MDL Drug Data Report
mHealth	Mobile health
MIC	Minimum inhibitory concentration
MIP	Molecular interaction profile
MMP	Matrix metalloproteinase
MS	Mass spectrometry
mTOR	Mammalian target of rapamycin
MW	Molecular weight
NCI	National Cancer Institute
NET	Norepinephrine transporter
NIH	National Institute of Health
NP	Natural products
NSCLC	Non-small cell lung cancer
PCR	Polymerase chain reaction

PDB	Protein Data Bank
PDGFR	Platelet-derived growth factor receptor
PNN	Probabilistic neural network
QSAR	Quantitative structure–activity relationship
SE	Sensitivity
SERT	Serotonin transporter
SNOMED	Systematized nomenclature of medicine
SP	Specificity
SRC	Proto-oncogene tyrosine-protein kinase Src
SVM	Support vector machine
TN	True negatives
TP	True positives
TTD	Therapeutic Target Database
UMLS	Unified medical language system
VEGFR	Vascular Endothelial Growth Factor Receptor
VS	Virtual screening
WHO	World Health Organization

List of Publications

1. Therapeutic Target Database Update 2014: a Resource for Targeted Therapeutics. **C. Qin**, C. Zhang, F. Zhu, F. Xu, S.Y. Chen, P. Zhang, Y.H. Li, S.Y. Yang, Y.Q. Wei, L. Tao and Y.Z. Chen. **Nucleic Acids Res.** 42(1):D1118-23 (2014).
2. What does it Take to Synergistically Combine Sub-potent Natural Products into Drug-level Potent Combinations? **C. Qin**, K.L. Tan, C.L. Zhang, C.Y. Tan, Y.Z. Chen and Y.Y. Jiang. **PLoS ONE**. 7(11):e49969 (2012).
3. Are Molecular Biomarker Based Mobile Health Technologies Ready for Healthcare Applications? **C. Qin**, L. Tao, C. Zhang, S. Y. Chen, P. Zhang, S. Y. Yang, Y. Q. Wei and Y. Z. Chen. **Submitted**.
4. A Resource for Facilitating the Development of Tools in the Education and Implementation of Genomics-Informed Personalized Medicine. C. Zhang, **C. Qin**, L. Tao, F. Zhu, S.Y. Chen, P. Zhang, S.Y. Yang, Y. Q. Wei, Y.Z. Chen. **Clin Pharmacol Ther.** 95(6):590-1. (2014).
5. Clustered Patterns of Species Origins of Nature-derived Drugs and Clues for Future Bioprospecting. F. Zhu, **C. Qin**, L. Tao, X. Liu, Z. Shi, X.H. Ma, J. Jia,

- Y. Tan, C. Cui, J.S. Lin, C.Y. Tan, Y.Y. Jiang and Y.Z. Chen. **PNAS**. 108(31):12943-8 (2011).
6. Nature's Contribution to Today's Pharmacopeia. L. Tao, F. Zhu, **C. Qin**, C.Zhang, F. Xu, C.Y. Tan, Y.Y. Jiang, Y.Z. Chen. **Nat Biotechnol**. Accepted (2014)
7. Therapeutic Target Database Update 2012: A Resource for Facilitating Target-Oriented Drug Discovery. F. Zhu, Z. Shi, **C. Qin**, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X.H. Liu, J.X. Zhang, B.C. Han, P. Zhang and Y.Z. Chen. **Nucleic Acids Res**. 40(D1):D1128-D1136 (2012).
8. Drug Discovery Prospect from Untapped Species: Indications from Approved Natural Product Drugs. F. Zhu, X.H. Ma, **C. Qin**, L. Tao, X. Liu, Z. Shi, C.L. Zhang, C.Y. Tan, Y.Y. Jiang and Y.Z. Chen. **PLoS ONE**. 7(7):e39782 (2012).
9. Combinatorial Support Vector Machines Approach for Virtual Screening of Selective Multi-Target Serotonin Reuptake Inhibitors from Large Compound Libraries. Z. Shi, X.H. Ma, **C. Qin**, J. Jia, Y.Y. Jiang, C.Y. Tan, Y.Z. Chen. **J Mol Graph Model**. 32:49-66 (2012).

Chapter 1: Introduction

1.1 Overview of targeted therapeutics in modern drug discovery

From ancient mysterious herbs to modern synthetic chemicals, drugs have been an integral part in people's health and well-being. It is for the benefit of the whole society to discover new drugs in the hope of defeating diseases and guarding health.

Because of the natural high demand for new drugs, abundant economic opportunities exist in the field of pharmaceutical industry. Especially in recent years, with the rapid development of biological technologies and huge advance in combinatorial chemistry, the drug discovery in pharmaceutical industry has received roaring attention and showed promising future. A tremendous amount of money, time and human resources have been injected into drug discovery, in the hope of finding new drugs. Statistics show that R&D expenditures in pharmaceutical industry has been growing at an annual growth of 13% since 1970, which leads to a total 50-fold increase and reaches 13% of the revenues of pharmaceutical companies. (1) And it is estimated that it would take 12-15 years, one billion US dollars on average in order to discover a new drug. (2)

A review of modern drug design stages will help to understand the lengthy and costly process to bring a new drug into market. As illustrated in Figure 1.1, the modern rational drug discovery starts with the identification of drug targets, by modulating which may result in desirable therapeutic effects to cure diseases or alleviate pain.

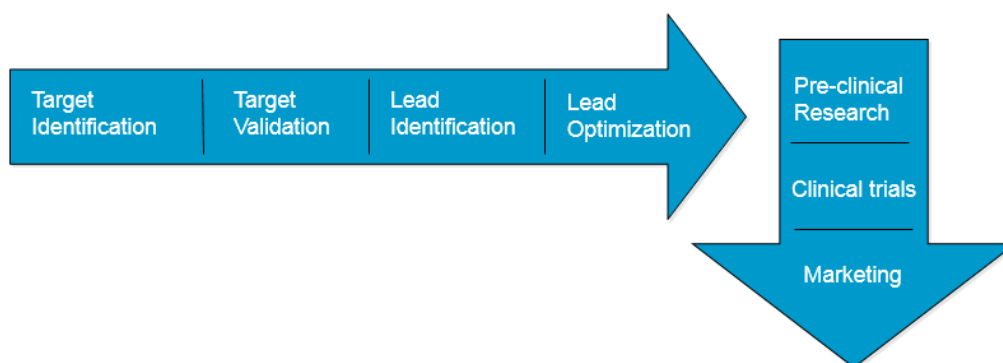


Figure 1. 1 Modern drug discovery process

This rational approach is in great contrast with traditional drug discovery, which mainly relies on “trial and error” approach or serendipitous discovery. Thanks to the advance in analytical chemistry and improvement of purification techniques in the early 20th century, active ingredients from traditional remedies and medicinal plants were characterized and some were extracted as drugs. Historically in classical pharmacology, collection of natural products, extracts and synthetic small molecules were screened against cell lines or organisms so as to identify the therapeutic chemicals. Later, with the advance of biochemistry that expanded people’s

understanding of diseases at the molecular level, it became clear that the therapeutic effects of many drugs came from their interactions with the biological molecules such as nucleic acids and proteins (like receptors, enzymes, ion channels, structural and transport proteins). And the modulation of such biological molecules, known as drug targets, can then result in desirable therapeutic effects.

Hence, the modern rational drug discovery is based on the hypothesis that modulation of certain targets may exert therapeutic effects and drugs directed at these targets are developed subsequently. A deep understanding of disease mechanism and molecular players within the disease pathology is required for target identification, but it is not sufficient to just identify the possible targets. Target validation is an important follow-on step to make certain that a potential target indeed plays a critical role in disease.

The purpose of target validation is to confirm the functions of potential targets and compounds modulating those targets will lead to therapeutic effects (3, 4). The target needs to be validated in various cell-based and animal disease models to establish and confirm its essential role in disease. (3, 5). Genetic methods, such as gene knockout and RNA interference, are commonly applied to conduct *in vivo* target validation (5, 6).

After target discovery, the following step of lead discovery starts with lead identification.(7, 8) A lead compound is a potentially therapeutically useful chemical which has pharmacological activity but requires structure modifications in order to become drug.(9) An assay to test the potency of compounds against the target needs to be developed and large chemical libraries of small molecules can then be screened using the developed assay. With the advance in automation systems, high throughput screening techniques have been applied extensively to rapidly test the activities of chemicals against the target. The emergence of combinatorial chemistry renders it possible to conduct systematic screening of a large number of small molecules with the maximized structure diversity. (10-12)

The leads identified from screening processes needs further evaluation on their potency, off-target activities and physiochemical or metabolic properties. Certain compound clusters with good pharmacological properties will be selected for modification before going into pre-clinical research stage when experiments on animal models are done to test the safety and efficacy profiles.

If everything goes well, the lead compounds will finally enter clinical trials and be tested on human beings. In recent years, the drug discovery process normally lasts for 12-15 years and the total discovery and development cost of a new drug is estimated to be as enormous as one billion US dollars.(2) Around half of the drug

development time and nearly two thirds of the cost needed for a new drug to gain FDA approval are devoted to clinical trials. (13) Overall, the phase I, phase II and phase III clinical trials are the most costly and time-consuming steps in the drug development processes. And the success of clinical trials relies heavily on the careful selection and strategic modification of lead compounds from previous steps.

In sum, the modern drug discovery, reliant on the biological insights of diseases, starts with the identification and validation of targets and then therapeutics directed to the targets are screened, optimized and selected to enter clinical trials. The targeted therapeutics are usually rationally designed drugs with promising efficacy profiles and few toxic effects. This is particularly true for cancer treatment. In the past, the standard treatment for cancer was the non-specific cytotoxic chemotherapy, which worked primarily through the inhibition of cell division and killed rapidly dividing cells in both cancer cells and human normal tissues. The past two decades see a dramatic shift of cancer drug development focus from the traditional cytotoxic chemotherapy to molecularly targeted therapeutics. The mechanisms of action and toxicity profiles of molecularly targeted drugs are different from traditional cytotoxic chemotherapy. The targeted cancer therapeutics interfere with specific molecular targets essential for tumor development and growth. Because such molecular targets are usually overexpressed or mutated in the cancer cells only, the targeted therapeutics are generally better tolerated in cancer patients than the

traditional cytotoxic chemotherapy. (14, 15)

1.2 The importance of multi-target therapeutics

However, a large percentage of drugs in development, which are typically directed at an individual target, sometimes show reduced efficacies and undesired safety and resistance profiles. For multigenic diseases, such as cancer, or diseases that act on different tissues or cell types, multi-target therapeutics can be more efficacious and less prone to resistance, compared to those drugs designed to act against an individual molecular target (16). Multi-target drugs, which are single chemical entities that act simultaneously at multiple molecular targets, and drug combinations, which are formulations of multiple active ingredients mixed in a single dosage form, are multi-target therapeutics studied in this thesis.

Multi-target therapeutics against selected multiple targets can selectively modulate the elements of those countertarget and toxicity activities, thus achieving enhanced therapeutic efficacies and improved safety and resistance profile. In particular, multi-target agents are able to regulate network robustness(17), redundancy(18), crosstalk(19), compensatory and neutralizing actions(20), anti-target and counter-target activities(21), and on-target and off-target toxicities(22).

Multi-target drugs tend to be more sparsely distributed in the chemical space than individual-target agents. For instance, known dual kinase inhibitors of selected kinase pairs are typically 10-fold smaller in numbers than the known individual kinase inhibitors (23). Therefore, exploration of larger chemical libraries may be needed for discovering new multi-target hits, particularly novel ones, against selected targets. To facilitate drug discovery, efficient methods to predict multi-target agents are highly desired.

Drug combinations are already standard treatments of many diseases including cancer, diabetes, viral and bacterial infections. And the high efficacy of existing drug combinations shows that searches to identify multi-target mechanisms can shed new light in drug discovery. Bioinformatics approaches to collect relevant drug combination data from literature and to analyse the pathways and molecular interaction profiles involved in drug combinations are expected to be useful.

1.3 More personalized targeted therapeutics driven by biomarkers

Modern drug development has been primarily focused on targeted therapeutics. (24-26) In recent year, there is an increasing movement towards stratified and personalized medicines.(27-29) Extensive efforts from the research, industry,

clinical, regulatory and management communities and the chemistry, biology, pharmaceuticals, and medicine disciplines have been collectively directed at the discovery, investigation, application, monitoring and management of targeted therapeutics and the diagnostic and prognostic biomarkers.(27, 30-33)

According to National Institute of Health (NIH), a biomarker is defined to be “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to therapeutic intervention”.(34) It can be a naturally occurring molecule, gene, or characteristic, which marks a physiological process or disease. Biomarkers can be used to support new medical diagnostics, preventive medicine, and drug development. The major disease related biomarkers are used to recognize an overt disease (diagnostic biomarkers) and to predict the disease progression (prognostic biomarker). Most relevant to drug discovery are the drug related theragnostic biomarkers that serve to guide treatment in various diseases. These theragnostic biomarkers can indicate the optimal dosage for a particular patient, or how the drug is metabolized in the body , or if the patient will be responsive to a certain drug treatment.

The clinical trial phases in modern drug discovery usually adopt a “one size fits all” approach and try to define the best treatment for the average patient in the whole

population. But it has been argued that such an approach may not be the most effective method for drug discovery.(35) The best treatment for an average patient may not be the optimal treatment option for a particular patient due to the heterogeneity of their molecular makeups among the patients. The use of biomarkers early in the drug development phase to select the likely responsive patients may help lower the attrition rate of clinical trials that results from patient heterogeneity in complex diseases like cancer. Only patients that are predicted to be responsive to a certain drug treatment will receive the drug in the clinical trials, thus decreasing the toxic effects and costs associated with the patients receiving ineffective treatments. As a result, the stratification of patients by companion biomarkers of a drug treatment will increase the success rate of clinical trials, accelerate the drug approval and provide the personalized treatment options.

In sum, the knowledge of the biomarkers will be useful not only for the discovery and development of targeted therapeutics and biomarkers (36, 37), but also for facilitating the development and practice of the diagnosis, prescription, monitoring and management of patient care in stratified and personalized medicines (27, 38, 39).

1.4 Bioinformatics methods for analysis of targeted therapeutics

The unprecedented development of technological advances and rapid accumulation of knowledge and information in biology and medicine through all fields of the “-omics” and translational researches signal the big data era of pharmaceutical information. The pharmaceutical industry is in dire need of an open-source, easily accessible, reliable and well-integrated data source. The large and complex collection of various targeted therapeutics data require effort to construct databases that can store, integrate and retrieve reliable information. Moreover, the known targeted therapeutics data in literature also call for bioinformatics analysis methods, which are of importance and practical usage to drug discovery.

1.4.1 The update of therapeutic target database to serve as an integrated information platform of targeted therapeutics

Drug discovery efforts can be facilitated by the information of drugs, targets, multi-target agents, drug combinations and biomarkers. Hence, to construct a database that provide such information will be a meaningful attempt.

Therapeutic Target Database (TTD, <http://bidd.nus.edu.sg/group/ttd/ttd.asp>) has been developed to provide comprehensive information about efficacy targets and the

corresponding approved, clinical trial and investigational drugs. The information of therapeutic targets and the targeted therapeutics included in TTD will facilitate target and drug discovery.

The aim of updating TTD is to make it into a more useful target and drug discovery resource in complement to other related databases such as drugbank (40). Continuous efforts have been made to provide the latest and comprehensive information about the targets, drugs and other therapeutics in different development and clinical stages, which is highly useful for focused drug discovery efforts and pharmaceutical investigations against the most relevant and proven targets. (24, 41) In addition to the update of these databases by expanded target and drug data contents, the usefulness of these databases for facilitating drug discovery efforts has been further enhanced by adding additional information and knowledge derived from the target and drug discovery processes.

Besides being a useful tool to store, retrieve and organize data, TTD serves an essential information platform for analyzing the data. Valuable information contained in the known drugs, targets, multi-target agents, drug combinations and biomarkers can be learnt through various bioinformatics analysis approaches. Many problems limiting drug discovery development can be addressed by such analysis and information learnt from the data in TTD will facilitate targeted therapeutics. The

procedures of this work will be the foundation of this thesis and will be discussed in detail in chapter 2.

To facilitate the modern drug discovery, various bioinformatics methods to analyze the targeted therapeutics based on the information from TTD have been developed and will be introduced in the following sections.

1.4.2 Machine learning methods to predict multi-target agents from large chemical libraries

The importance of multi-target agents for treatment of complex diseases and the difficulty to predict them due to their limited number have been introduced in chapter 1.2. In this thesis an efficient virtual screening method to predict multi-target agents will be presented to address the problem. It is therefore necessary to give an overview of virtual screening.

In the lead discovery of rational drug design process, although the high throughput screening techniques significantly reduce the time and cost consumed for screening an individual compound (28), the daunting size of combinational chemical library is beyond the capability of wet-lab experimentations.

To solve the problem, various virtual screening techniques have been developed in recent years to help accelerate the lead discovery and make it more efficient (18-24). Two different strategies are used in virtual screening, namely structure-based and ligand-based.

If the 3D structure of a validated therapeutic target is known, a structure-based virtual screening technique such as docking can be used to estimate the possibility of good binding affinity between the compound and the target. Through molecular modeling calculations, a scoring function could be applied to give a score to indicate the possibility.

Given a set of compounds with known structures that bind to the target, a ligand-based virtual screening approach could be used. Chemical structure similarity search, pharmacophore models that capture the information shared by the known ligands, machine learning models to learn from known structures and generate rules to describe substructure features or physiochemical features, quantitative structure activity relationship models that correlate the activities and quantitative structure properties of ligands are popular approaches in ligand-based virtual screening.

In-silico methods mentioned above have been extensively explored as virtual screening (VS) tools for facilitating the discovery of multi-target agents from both

focused and large chemical libraries.(23, 42-47) One popular strategy in searching multi-target agents is to combinatorially use individual-target VS tools to separately screen chemical libraries against each of a group of selected multiple targets for finding virtual hits active against all of the selected targets.(45)

The level of success in screening larger chemical libraries depends on the ability of VS tools to produce sufficiently high hit retrieval rates (yields) and low false hit rates. (23, 45) High yields against individual target compensate for the reduced collective yields against multiple targets (if the yield for one target is 50~70%, the collective yield for two targets may be statistically reduced to 25~49%). Low false hit rates ensure that virtual hits are sufficiently enriched with true multi-target agents (i.e. sufficient percentage of virtual hits are true hits). Multi-target VS performance may also be affected by the similarity level of the drug-binding domains of the selected multiple targets. For multiple targets of higher similarity levels, it may be harder to distinguish multi-target agents from individual-target agents, because of the smaller differences between their structures. Likewise, inhibitors of other similar targets may also be falsely recognized as multi-target agents of selected multiple targets because of binding site similarity between the other similar target and one or more of the multiple targets of the multi-target agents.

Therefore, VS methods for screening multi-target agents need to be rigorously

evaluated not only on the yield of multi-target hits but also on target selectivity (ability to distinguish multi-target agents from individual-target agents and agents of similar targets). Ligand based machine learning methods such as combinatorial support vector machines (Combi-SVM) have recently been explored as VS tools for searching dual kinase inhibitors and dual target serotonin reuptake inhibitors from large compound libraries.(23) The machine learning methods work well in predicting serotonin reuptake inhibitors with fair yields, moderate to good target selectivity and low false hit rates. More comprehensive tests against diverse sets of target pairs of different biochemical classes and varying similarity levels are needed to fully evaluate the potential of these and other VS methods in searching multi-target agents.

1.4.3 Clustering method to analyse the distribution patterns of targeted drugs in target-specific chemical space

Using machine learning methods, virtual multi-target hits could be identified from large chemical databases. A small percentage of virtual hits validated by wet-lab experiments will enter the next phase of drug discovery. From hit to lead to clinical trial drugs and to approved drugs, the drug discovery process is still time and money consuming (48) and methods to identify drug candidates are desired.

Out of the many compounds that pass virtual screening and subsequent experimental

validation to be identified as leads, few of them could pass clinical trial phase and enter the market. According to a number of reports (1, 49) , the drug failure rate during clinical trials have increased significantly, with an estimation as high as 90% (50).

Myriad of reasons are behind the high drug failure rate. Problems such as poor bioavailability (39% of failures), low clinical efficacy (30% of failures), toxicity (11%) and adverse reactions in humans (10%) have been reported as the major causes for failure.(50) Other than these, commercial reasons, formulation issues and so on also contribute to the high attrition rate. (51)

All these problems become more frequent in recent years, mostly because of the advent of combinatorial chemistry. The evolutionary invention of high-throughput screening (HTS) has generated a very large number of potential drugs. Traditionally, the most important criterion for selecting the compounds to synthesize for HTS was to increase the chemical diversity of the structure. However, other qualities that finally influence whether a compound would pass the clinical trials, like efficacy, toxicity, bioavailability and so on, were largely ignored.(52) Hence, since the introduction of combinatorial chemistry, this neglect has led to a shift in those quality profiles of compound libraries available for drug development, which in turn negatively affects the efficacy of HTS screened compounds that will proceed into

the clinical phases.(53)

More clues to drugs can be obtained from the distribution profiles of approved and clinical-trial drugs. Using the clustering method, the large chemical space could be explored and areas where drugs are more likely to come from could be found.

Restricted by the structure conformation of targets, hits of a certain target generally adopt several specific scaffolds. The approved and clinical trial drugs are composed of a limited number of molecular scaffolds (54-56) in contrast to the high number of bioactive molecular scaffolds (57, 58). For instance, many drugs have been derived from individual scaffold groups such as macrocycles (59), and 12 FDA approved anticancer kinase inhibitor drugs (60, 61) are grouped into three scaffold groups (62). Investigation and exploration of these highly privileged drug scaffolds are important for discovering new drug-like scaffolds, molecular analogs and new drugs. From hits to leads to final clinical phases, the potential for drug development of a hit depends not only on its bioactivity, but also on the properties of its structures, such as optimized drug-likeness(63) and minimized unwanted properties (64, 65).

Questions arise as to whether drugs of a target tend to cluster together in the chemical space and whether the scaffolds of hits in those drug-clustered areas are more likely to be drug productive. Through clustering drugs, inhibitors and similar

compounds of a certain target, the above questions could be investigated and possibly answered.

In sum, we expect that insights obtained from clustering patterns will give more clues to the drug development. But due to the limit in time and computational power and the huge amount of data to cluster in order to analyze the patterns, only preliminary results will be outlined in chapter 3 part 2.

1.4.4 Systematic analysis to study synergistic combinations of natural products as potential sources of multi-targeted therapeutics

Partly because of low drug productivity from high-throughput screening and combinatorial chemistry based drug discovery programs, there have been renewed interests in the exploration of natural products (NP) as sources of new drugs (60-62, 66-69). In particular, NP combinations, in many cases as combinations of whole herbs or herbal extracts, have been extensively studied (70, 71), tested in clinical trials (72-74), and widely used in traditional, folk and alternative medicines (75, 76). These NP combinations may be useful sources for developing new drug combinations based on their novel multi-targeted mechanisms (16, 73, 77) or molecular scaffolds (58) to meet the increasing demand for multi-targeted drug

combinations (78).

Opinions vary regarding to the therapeutic efficacies of NP combinations. One attributes the efficacies of NP combinations to placebo effects (79-81) based on some indications from clinical trials (80, 81) and the findings that many bioactive NPs are sub-potent with respect to drugs (82, 83). Another credits the efficacies of NP combinations to synergistic effects (71, 73, 82, 84-86) based on the findings that some NP combinations produce significantly better effects than equivalent doses of their components (82, 85) and clinical outcomes are not necessarily influenced by positive beliefs (79).

The contribution of synergistic effects to therapeutic efficacies has been extensively studied (71, 73, 85, 86). While many studies have consistently suggested that therapeutic potency can be enhanced by synergistic effects, the levels of potency enhancement, particularly with respect to those of drugs, have not been systematically studied to quantitatively assess the contribution of synergism to the therapeutic efficacies of NP combinations.

The potency difference between natural products and drugs, the feasibility and molecular basis to recover the difference by synergistic combination will be addressed in chapter 4.

1.4.5 Analysis of biomarker for personalized medicine

The information of targeted therapeutics and biomarkers may be potentially incorporated into the widely-used disease classification systems for more refined classification of disease subclasses and patient subpopulations responsive to a particular treatment so as to better facilitate the diagnosis, prescription, monitoring and management of patient care in stratified and personalized medicines. While the information about targeted therapeutics and biomarkers can be obtained from the established drug (87), efficacy target (88) and biomarker (89-91) databases, the data access modes of these database are not specifically designed for optimally supporting such tasks. There is a need to introduce new access modes based on such widely-used disease classification systems as the International Classification of Diseases (ICD) codes developed by the World Health Organization (WHO) (92, 93). These new access modes also enable broader, more convenient and automatic data access, processing and exchange by all bench-to-clinic communities particularly non-domain experts. By analyzing TTD biomarker information, this new access mode will be elaborated in chapter 4.1.

A high level of interest in mobile health (mhealth) has emerged recently, as exemplified by the US Secretary of Health and Human Services Kathleen Sebelius'

reference of mHealth as “the biggest technology breakthrough of our time” and its use as being able to “address our greatest national challenge.” A Pubmed keyword search using “mhealth” showed 245 publications in 2012-2014 compared to only 41 publications before 2012.

Increasing efforts have been directed at the development of molecular biomarkers and the new detection technologies into mhealth devices to extend the coverage and improve quality of the mhealth. But there are questions about whether molecular biomarkers combined with the new technologies are ready for mhealth applications: (1) whether the new technologies are sufficiently sensitive, fast and inexpensive for biomarker detection, (2) the relevance and accuracies of the literature-reported non-invasive molecular biomarkers for mhealth applications, (3) how the healthcare providers cope with the increased workload resulting from widespread use of mhealth devices.

These questions can be addressed by analyzing the literature-reported biomarker detection capability (detection sensitivity, required sample volume, testing time, and cost) of the new technologies, and the relevance (disease coverage, patient populations) and diagnostic/prognostic accuracies of the 664 literature-reported non-invasive molecular biomarkers stored in TTD for mhealth applications. As a byproduct, the feasibility and potential issues of workload reduction by developing

and using a digitally-coded biomarker, disease and therapeutic information processing system for electronically pre-screening the mhealth biomarker readings will be discussed in chapter 4.2.

1.5: Outline of thesis

The main body of this thesis starts with the update of Therapeutic Target Database (TTD) in Chapter 2 as a meaningful effort to curate, store, integrate and retrieve data of various types of targeted therapeutics, including drug targets, drug chemicals, natural products and biomarkers. The drug and target information stored in TTD was updated constantly to include recent approvals into clinical trials and markets, new categories of information such as multi-target drugs and drug combinations were added to TTD in the last update and the most recent update incorporated the biomarker information and linking of all data through international classification of disease code. Through these updates, TTD serves as a comprehensive and integrated information source of targeted therapeutics to facilitate drug discovery. And various bioinformatics methods based on data from TTD were developed and discussed in subsequent chapters.

Chapter 3 describes several methods to facilitate the design of traditional

multi-target small molecule drugs. Three machine learning methods, support vector machine (SVM), K-Nearest Neighbor(kNN) and probabilistic neural network (PNN) were implemented as virtual screening tools to predict dual target inhibitors from large chemical libraries such as MDDR and pubchem. Models of 29 targets pairs with varying similarity levels between their drug-binding domains were developed. And the multi-target hit and target selection performances of the combinatorial SVM, KNN and PNN were evaluated in detail in Chapter 3.1.

Using machine learning methods, virtual hits could be identified, but from hit to lead and from lead to drugs, methods were still in demand to identify compounds of good structure scaffold and optimal drug property that could have higher chance to enter clinical trials and become drugs. In chapter 3.2, a hierarchical clustering method was developed to cluster drugs, inhibitors of a specific target in the chemical space spanned by structurally similar bioactive and non-bioactive compounds. Preliminary investigation of the plausible drug distribution patterns in the chemical space is outlined in hope to give more clues to drug development.

Partly due to the low productivity of virtual screening and synthetic chemistry, interests in natural product has been renewed. In particular, the natural product combinations may be useful sources for developing new drug combinations based on their novel multi-target mechanisms or molecular scaffolds. In chapter 4, a

systematic analysis of synergistic natural product combinations was described. The potency difference between natural products and drugs, the feasibility and molecular basis to recover the difference through synergistic combinations are addressed and specific multi-target modes are identified.

Chapter 5 is devoted to reflect the current shift of drug development focus to more personalised targeted therapeutics. Current biomarkers in TTD will be analyzed with respect to disease subtype classifications. More refined classification of patient subpopulations for personalized targeted therapeutics will be proposed. In addition, the feasibility of utilizing non-invasive biomarkers for mHealth applications are analyzed and discussed.

The last chapter 6 summarises all the major findings and merits from the research works described in the previous chapters and future works to further develop the targeted therapeutics are described.

Chapter 2 Update of therapeutic target database as an integrated source of targeted therapeutics data

Therapeutic Target Database, first developed in 2002, has been in the frontier to provide reliable information about therapeutic targets. In the past decade with rapid progress in target discovery, TTD still remained one of the most popular and mostly accessible database to provide pharmaceutical information on therapeutic targets. Drugs that act on novel targets have been approved or entered clinical trials, and new pharmaceutical information regarding drugs and therapeutic targets have emerged. To keep the data in pace with the current drug discovery progress, it is of great importance to update the data in TTD. And not only should information regarding drugs and drug targets be updated, but also other relevant information about other therapeutics such as multi-target agents, natural products and biomarkers should be incorporated in the database.

In this chapter, the details of the updates to TTD will be explained, with the focus on the collection and access of data. And the data in TTD act as the foundation of the work described in the subsequent chapters. Hence, various aspects of data in TTD will be elaborated, but the construction of TTD database such as design and implementation of this relational database will be skipped from this thesis, as the update of TTD was a collective effort and my work was mainly on the collection and curation of data.

2.1 Statistics of updated targeted therapeutics in TTD

Major improvements were made to the Therapeutic Target Database (TTD, <http://bidd.nus.edu.sg/group/ttd/ttd.asp>) to facilitate target-oriented drug discovery in the past two updates (2012 update and 2014 update).

As a popular publicly accessible database that provides comprehensive information of targets and drugs, the target and drug data in TTD were significantly expanded to its current status of 388 successful, 461 clinical trial, and 1,467 research targets; 2003 approved (1,008 nature product derived), 3,147 clinical trial, 498 discontinued clinical trial and 14,856 experimental drugs. These are compared to the 364 successful, 286 clinical trial, and 1,331 research targets; 1,540 approved (939 natural product derived), 1,423 clinical trial, 345 discontinued clinical trial and 14,853 experimental drugs in the 2012 update of TTD and the 348 successful, 249 clinical trial, 43 discontinued clinical trial and 1254 research targets, and 1514 approved, 1212 clinical trial and 2302 experimental drugs. Newly approved drugs and targets were constantly kept track of and deposited into the database, as well as novel drug molecules that recently entered the clinical trials. There was a tremendous increase of the number of experimental drugs included in TTD over the past few years so as to make TTD a comprehensive source of drug information.

Other than drugs and targets, multi-target agents, drug combinations and synergistic natural product combinations were added and significantly expanded in order to facilitate target-oriented drug discovery. Currently, 20,818 multi-target agents against 385 targets pairs and 115 drug combinations are collected in the TTD, in comparison to 3,681 multi-target agents in 2012 update and zero drug combinations in 2010 update. The incorporation of these multi-target therapeutics currently in development or sold in the market would make TTD more useful for researchers studying complex heterogenic diseases.

Target validation data such as the experimentally measured potency of 11,810 drugs against 915 targets, the observed potency or effects of 1,436 drugs against 274 cell-lines and 497 drugs against disease models (*ex vivo*, *in vivo* models), and the observed effects of target knockout, knockdown or genetic variations for 307 targets were included in TTD.

The major improvement of latest TTD update was the incorporation of the information of 1,755 biomarkers for 365 disease conditions to better serve the multiple bench-to-clinic communities and to facilitate the development and practice of stratified and personalized medicines. And this feature will be further elaborated in detail in the following sections.

The statistics of our updated data is summarized in **Table 2.1**.

Table 2. 1 Statistics of the drug targets, drugs and their structure and potency data in 2014 version of TTD database.

		2014	2012
		Update	Update
Statistics of Drug Targets	Number of All Targets	2,360	2,025
	Number of Successful Targets	388	364
	Number of Clinical Trial Targets	461	286
	Number of Research Targets	1,467	1,331
Statistics of Drugs	Number of All Drugs	20,667	17,816
	Number of Approved Drugs	2,003(1,008)	1,540
	(No of Natural Product Derived		(939)
	Drugs)		
	Number of Clinical Trial Drugs	3,147(369)	1,423
	(No of Natural Product Derived		(369)
	Drugs)		
	Number of Discontinued Drugs	498	345
	Number of Pre-Clinical Drugs	163	165
	Number of Experimental Drugs	14,856	14,853
	Number of Multi-Target Agents	20818	3,681

	Number of Drug Combinations	115	115
Statistics of Drugs with	Number of Small Molecular Drugs	17012	14,170
Available Structure or	with Available Structure		
Sequence Data	Number of Antisense Drugs with	652	652
	Available Sequence Data		
	Number of Agents with Potency Data	11810	11,810
	Against Target		
Statistics of Drugs with	Number of Agents with Potency Data	1753	497
Activity Data or	Against a Disease Model Such As a		
Structure-Activity	Cell-line, ex-vivo, in-vivo Model		
Relationship	Number of Quantitative		
	Structure-Activity Relationship QSAR	841 (228)	841
	models		(228)
	(No of Chemical Types)		

2.2 Materials and methods.

2.2.1 Data collection method

The relevant information of interest is scattered in the vast collection of medical and biological literature such as research articles, reputable review journals, conference proceedings and specialized books. Depending on the source of literature and article

types, the methods to extract information and to collect data vary. Because of the huge amount of literature to search from, an automated literature information extraction workflow is desirable and needs to be developed. But it is generally difficult to use automatic text mining methods to recognize terms from different sources due to the abbreviations, synonyms and different expressions. To ensure the reliability of collected data, manual search and curations are necessary. Hence, in the development of TTD, the automatic text mining techniques as well as manual searching were both used and complemented each other.

In house perl scripts were developed to automatically screen and extract literature containing relevant keywords or specific word patterns. The matched literature was then examined manually to search for desired information. The automatic text search was generally done on abstracts of research articles or conference proceedings first. Sometimes meaningful information could already be identified from the abstracts, but in many other cases, full articles must be downloaded and read through carefully for reliable data and details.

In addition to literature, many established online databases provided valuable information and served as curation tools. Chemical databases such as Pubchem, MDDR, ChEMBL, bindingDB, drugbank and Zinc, biological databases such as swissprot, uniprot, PDB and KEGG are useful sources of information. In general,

SQL query languages were used to automatically extract various kinds of information from databases and perl scripts were written to further process and format the extracted data. A significant portion of information from different databases were shared, but in different formats or presented differently. Efforts were done to integrate data from various sources. Through comparison of similar data from different sources, the quality of the shared data could be guaranteed. When conflicts arise, manual examination is done to further ensure the quantity of data by searching for original literature.

2.2.2. Data sources

Following the general information mining method presented above, the sources of various information in TTD will be described in detail.

Newly approved drugs and therapeutic targets were collected from FDA Drugs@FDA database (<http://www.accessdata.fda.gov/scripts/cder/drugsatfda/>) and comprehensive search of literatures. Information of drugs and targets in clinical trial development was obtained from the latest reports of pharmaceutical companies namely Astrazeneca, Bayer, Boehringer Ingelheim, Genentech, GSK, Indenix, Incyte, ISIS, Merck, Novartis, Pfizer, Roche, Sanofi Avetis, Schering-Plough, Spectrum, Takeda and Teva. The literature search was done by searching the

Pubmed using keywords “drug” and “target”, “therapeutic” and “target”, “clinical trial” and “drug”, “clinical trial” and “target”. In addition, some of the newly added clinical trial drug information came from abstract of American Society of Clinical Oncology annual meeting. From 1995 to 2011, 2665 abstracts were evaluated manually to extract information regarding the clinical trial drugs, such as phase, number of participants, result, endpoint, year of study and study design. The updated information from clinicaltrial.gov database were also incorporated in the database.

Experimentally determined cell-based inhibitory activities of anticancer and antibacterial natural products were searched from the Pubmed database (94) by using keyword combinations of ‘natural product’, ‘herb’, ‘medicinal plant’, ‘extract’, ‘ingredient’, ‘GI50’, ‘IC50’, ‘MIC’, “activity”, ‘cell-line’, and ‘in vitro’. Till now, cell-based inhibitory activities of 1378 anticancer and antimicrobial natural products and 99 antimicrobial natural product extracts were obtained from the literatures. For natural product with multiple potency data, the best potency was selected and stored.

Literature-reported synergistic natural product combinations were searched from the Pubmed database by using keyword combinations ‘natural product’, ‘herb’, ‘medicinal plant’, ‘extract’, ‘ingredient’, ‘synergistic’, ‘synergy’, ‘synergism’, ‘synergize’, and ‘potentiate’. The full reports of the searched articles were evaluated

to select those synergistic natural product combinations with the experimental cell-based activities available for all constituent natural products both as individual and in the respective combination.

For multi-target agents, Pubmed database was searched upon using such keywords as ‘multi-target’, ‘dual target’ and ‘dual inhibitor’. Multi-target agent against a target pair was defined to be a compound active against both targets at potency values of $\leq 20 \mu\text{M}$ regardless of their possible activities against other targets. The 3D structures of these multi-target agents were generated by using CORINA (39) from the 2D structures manually drawn based on the literature provided structures or the structures found in such chemical databases as BindingDB (40), ChEMBL (41) and PubChem (28).

To broadly cover various types of biomarkers, comprehensive literature search was conducted in Pubmed database by using combination of keywords “biomarker”, “clinical”, “patient”, “disease”, “drug”, and specific disease names. Over 100 review papers from reputable journals were downloaded and read through in detail.

Additional sources such as the abstracts of the American society of clinical oncology were also systematically searched. From 1995 to 2013, over 700 biomarker related abstracts were processed through data mining and curated manually. As far as

possible, original research papers cited in the review were checked to collect detailed information about the biomarker. And the details collected include the name of biomarker, type of biomarker, source of biomarker, measurement, detection method, detection threshold, specific disease, specific function of the biomarker, study design, number of participants in the study and result.

2.3. Data in TTD and ways to access them



2.3.1 Overall search and download options

Starting from the home page of TTD (**Figure 2.1**), the users can easily search the whole TTD database. Five search fields of disease, drug, drug target, biomarker and drug scaffold are listed to help users with different search needs. Customized keyword search is also possible in TTD, by accessing the customized search tab (**Figure 2.2**). Target name, drug name, disease indication, target biochemical class, drug mode of action and therapeutic class are the specific customized search fields for drug and target search. The customized search of biomarkers by their development status can be achieved by clicking on “Search biomarker” (**Figure 2.3**). Other than these search methods, target similarity search and drug similarity search to search for similar target and drugs given a target FASTA sequence or drug structure can be done as well.

If the users are not satisfied with the online search and would like to have access to more information, full TTD data download is also provided to facilitate batch processing of the various data types in TTD (**Figure 2.4**). Full target information, ID mapping of TTD data to public databases, synonyms of chemicals in TTD, drugs, target and biomarker mapping to diseases can be easily downloaded and analyzed. Specific target information such as sequence and uniprot ID and drug structures in SDF format are prepared for download to enable easy analysis and further processing of TTD data.

By inputting keywords in different fields, the users will get access to search pages containing all relevant information. And hyperlinks to target, drug and biomarker detail information pages are listed.

Therapeutic Targets Database



[HOME](#) [Customized Search](#) [Target Similarity Search](#) [Drug Similarity Search](#) [Download](#)
[QSAR Models](#) [Target Validation](#) [Multi Target Agents](#) [Drug Combinations](#) [Nature-derived Drugs](#)

Search Whole Database

Search drugs and targets by disease or ICD identifier: [ICD9 Index](#) [ICD10 Index](#)

Examples: Alzheimer; 331 or ICD9:331; G30 or ICD10:G30; ...

Search for drugs:

Examples: Oseltamivir; Alzheimer's disease; ...

Search for targets:

Examples: Muscarinic acetylcholine receptor; Non-small cell lung cancer; ...

Search for biomarkers: [ICD9 Index](#) [ICD10 Index](#)


Examples: p53; Alzheimer; 331 or ICD9:331; G30 or ICD10:G30; ...

Search for drug scaffolds:


Please Select a Drug Scaffold Name ▼

Read more about TTD [Query Methods](#)

Figure 2. 1 Screenshot of TTD home page.



Therapeutic Targets Database




[HOME](#)
[Customized Search](#)
[Target Similarity Search](#)
[Drug Similarity Search](#)
[Download](#)

[QSAR Models](#)
[Target Validation](#)
[Multi Target Agents](#)
[Drug Combinations](#)
[Nature-derived Drugs](#)

Field Name	Match Text
Target Name	<input style="width: 90%;" type="text"/> <div style="display: flex; justify-content: space-between; margin-top: 5px;"> <input checked="" type="radio"/> All <input type="radio"/> Successful <input type="radio"/> Clinical Trial <input type="radio"/> Research </div>
Drug Name	<input style="width: 90%;" type="text"/> <div style="display: flex; justify-content: space-between; margin-top: 5px;"> <input checked="" type="radio"/> All <input type="radio"/> Approved <input type="radio"/> Clinical Trial </div>
Disease Indication	<div style="border: 1px solid #ccc; padding: 2px;">Please Select a Disease Name ▼</div>
Target BioChemical Class	<div style="border: 1px solid #ccc; padding: 2px;">Please Select a Target BioChemical Class ▼</div>
Drug Mode of Action	<div style="border: 1px solid #ccc; padding: 2px;">Please Select a Drug Mode of Action ▼</div>
Drug Therapeutic Class	<div style="border: 1px solid #ccc; padding: 2px;">Please Select a Drug Therapeutic Class ▼</div>
<input type="button" value="Submit"/> <input type="button" value="Reset"/>	

Figure 2. 2 Screenshot of TTD customized search

Therapeutic Targets Database





[HOME](#) [Customized Search](#) [Target Similarity Search](#) [Drug Similarity Search](#) [Download](#)

[QSAR Models](#) [Target Validation](#) [Multi Target Agents](#) [Drug Combinations](#) [Nature-derived Drugs](#)

Search Targets and Drugs

Search Biomarkers

Field Name	Match Text
Biomarker	<input type="text"/>
ICD9 Index ICD10 Index	<div><input type="radio"/> All <input type="radio"/> Clinically Used <input type="radio"/> Clinical Trial <input checked="" type="radio"/> Reseearch</div> <div>Examples: p53; Alzheimer; ICD9:331; ICD10:G30; ...</div>
<div>SearchReset</div>	

Figure 2. 3 Screenshot of TTD customized search of biomarkers

HOME	Customized Search	Target Similarity Search	Drug Similarity Search	Download
QSAR Models	Target Validation	Multi Target Agents	Drug Combinations	Nature-derived Drugs
TTD Database Downloads				
Download TTD targets information in raw format				Click to Save
Cross-matching ID between TTD drugs and public databases				Click to Save
Synonyms of drugs and small molecules in TTD				Click to Save
Drug to disease mapping with ICD identifiers				Click to Save
Target to disease mapping with ICD identifiers				Click to Save
Biomarker to disease mapping with ICD identifiers				Click to Save
Target Information Downloads				
Download Uniprot IDs for all targets				Click to Save
=>Download Uniprot IDs for successful targets only				Click to Save
=>Download Uniprot IDs for clinical trial targets only				Click to Save
=>Download Uniprot IDs for research targets only				Click to Save
Download sequence data for all targets				Click to Save
=>Download sequence data for successful targets only				Click to Save
=>Download sequence data for clinical trial targets only				Click to Save
=>Download sequence data for research targets only				Click to Save
Drug Structure Downloads				
Download structure data for all drugs in SDF format				Click to Save
=>Download structure data for approved drugs only				Click to Save
=>Download structure data for clinical trial drugs only				Click to Save
=>Download structure data for experimental agents only				Click to Save
Download antisense oligonucleotide sequences in raw format				Click to Save
Last update by				

Figure 2. 4 Screenshot of database download page in TTD.

2.3.2 Targets and drugs

The detailed information page for each target includes target name, development status of target (successful, clinical trial and experimental), synonyms, disease, drugs directed at this target, biochemical class, EC number, pathways, Uniprot ID, PDB structure ID, sequence, target validation link, inhibitors of this target included in TTD, Multitarget drugs, cross links to 3D structure, related literature and online medical dictionary and literature references, if available. Screenshots using ABL1 protein target as an example can be referred to in Figure 2.5. Many of the fields contain hyperlinks to external databases such as swissprot, PDB, KEGG and pubmed databases for convenient retrieval of further information.

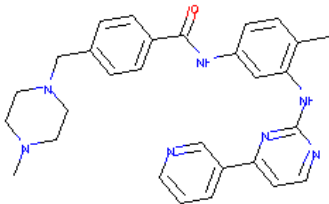
TTD Target ID: TTDS00346					
Target Information					
Name	B-Raf proto-oncogene serine/threonine-protein kinase				
Type of target	Successful target				
Synonyms	B-Raf				
	B-raf protein				
	BRAF				
	BRAF serine/threonine kinase				
	BRAF(V599E)				
	P94				
	V-Raf murine sarcoma viral oncogene homolog B1				
Disease	Malignant melanoma [ICD9: 172 ICD10: C43]				[1]
	Melanoma [ICD9: 172 ICD10: C43]				[2]
	Pancreatic cancer [ICD9: 157 ICD10: C25]				[3]
Drug(s)	Vemurafenib	Drug Info	Approved	BRAF-positive unresectable or metastatic melanoma	
	Sorafenib	Drug Info	Launched	Advanced renal cell carcinoma	[4][5][6]
	R7204	Drug Info	Phase III	Malignant melanoma	[7]
	RG7204	Drug Info	Phase III	Adjuvant metastatic melanoma, BRAF mutation positive	
	RG7421+RG7204	Drug Info	Phase III	Metastatic melanoma BRAF mutation positive	
	Sorafenib	Drug Info	Phase III	Hepatocellular carcinoma, NSCLC, melanoma	[4][5][6]
	Trametinib + dabrafenib	Drug Info	Phase III	Metastatic melanoma, adjuvant therapy	
	Vemurafenib	Drug Info	Phase III	Adjuvant metastatic melanoma, BRAF mutation positive	
	Dabrafenib	Drug Info	Phase II	Non-small cell lung cancer	
	PLX4032	Drug Info	Phase II	Throid cancer	
	RG7204	Drug Info	Phase II	Papillary thyroid cancer, BRAF mutation positive	
	Sorafenib	Drug Info	Phase II	Myelodyspalstic syndrome, AML, head & neck cancer, breast,	[4][5][6]

	Sorafenib	Drug Info	Phase II	myeloid leukemia, myelodysplastic syndrome, AML, head & neck cancer, breast, colon, ovarian, pancreatic cancer	[4][5][6]
	Trametinib + dabrafenib	Drug Info	Phase II	Colorectal cancer	
	Vemurafenib	Drug Info	Phase II	Papillary thyroid cancer, BRAF mutation positive	
	GSK2118436	Drug Info	Phase I/II	Metastatic melanoma, solid tumors	[8]
	ARQ 736	Drug Info	Phase I	Late-stage solid tumors	
	BMS-908662	Drug Info	Phase I	Cancers	[9]
	Dabrafenib	Drug Info	Submitted	Metastatic melanoma	
	RG7204	Drug Info	Filed	Metastatic melanoma, BRAF mutation positive	
	RG7204	Drug Info	Phase I	Metastatic melanoma, BRAF mutation positive	
	RG7256	Drug Info	Phase I	Malignant melanoma	
	Trametinib + dabrafenib	Drug Info	Submitted	Metastatic melanoma	
	Vemurafenib	Drug Info	Phase I	Metastatic melanoma, BRAF mutation positive	
BioChemical Class	Transferases transferring phosphorus-containing groups				
EC Number	EC 2.7.1.37				
Pathway	Acute myeloid leukemia				
	Bladder cancer				
	Chemokine signaling pathway				
	Chronic myeloid leukemia				
	Colorectal cancer				
	Endometrial cancer				
	ErbB signaling pathway				
	Focal adhesion				
	Glioma				
	Insulin signaling pathway				
	Long-term depression				
	Long-term potentiation				
	MAPK signaling pathway				
	Melanoma				
	Natural killer cell mediated cytotoxicity				
	Non-small cell lung cancer				

Pathway	Long-term potentiation		
	MAPK signaling pathway		
	Melanoma		
	Natural killer cell mediated cytotoxicity		
	Non-small cell lung cancer		
	Pancreatic cancer		
	Pathways in cancer		
	Prostate cancer		
	Regulation of actin cytoskeleton		
	Renal cell carcinoma		
	Thyroid cancer		
	Vascular smooth muscle contraction		
	mTOR signaling pathway		
UniProt ID	P15056		
PDB Structure	1UWH ; 1UWJ ; 2FB8 ; 2L05 ; 3C4C ; 3D4Q ; 3IDP ; 3II5 ; 3NY5 ; 3OG7 ; 3PPJ ; 3PPK ; 3PRE ; 3PRI ; 3PSB ; 3PSD ; 3Q4C ; 3Q96 ; 3SKC ; 3TV4 ; 3TV6 ; 4DBN ; 4E26 ; 4E4X ; 4EHE ; 4EHG ; 4FK3 ; 4G9C ; 4G9R ; 4H58 ; 4JVG .		
Function	Involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May play a role in the postsynaptic responses of hippocampal neuron.		
Sequence	MAALSGGGGGGAEPGQALFNGDMEFEAGAGAGAAAASSAADPAIPEEVWNIKQMIKLTQEH IEALLDKFGGEHNPPSIYLEAYEYTSKLDALQQREQQLLESLGNGTDFSVSSSASMDTV TSSSSSSLSVLPSSLSVFNPTDVARSNPKSPQKPIVRVFLPNKQRTVVPARCGVTVRDS LKKALMMRGLIPECCAVYRIQDGEKKPIGWDTDISWLTGEELHVEVLENVPLTTHNFVRK TFFTAFCDFCRKLQFGFRCTCGYKFHQRCSTEVPLMCVNYDQLDLLFVSKFFEHHPI PQEEASLAETALTSGSSPSAPASDSIGPQILTSPSPSKSIPIQPFRPADEDHRNQFGQR DRSSAPNVHINTIEPVNIDDLIRDQGGFRGDDGGSTTGLSATPPASLPGLTNTVKALQKSP GPQREKSSSSSEDRNRMKTLGRDSSDDWEIPDGQITVGQRIGSGSFGTVYKKGWHDV AVKMLNVTAPTQQQLQAFKNEVGVLKTRHVNILLFMGYSTKPQLAIVTQWCEGSSLYHH LHIETKFEMIKLIDARQTAQGM DYLHAKSIIHRDLKSNFIHEDLTVKIGDFGLATV KSRWSGSHQFEQLSGSILWMAPEVIRMQDKNPYSFQSDVYAFGIVLYELMTGQLPYSNIN NRDQIIFMVG RGYLSPDL SKVRSNCPKAMKRLMAECLKKRDERPLFPQILASTELLARS LPKIHRSAEPLNRAGFQTEDFSLYACASPKTPIQAGGYGAFFVH		
Target Validation	Click to Find Target Validation Information.		
Inhibitor	BMS-908662	Drug Info	[9]
	GSK2118436	Drug Info	[8]
	R7204	Drug Info	[7]
	Sorafenib	Drug Info	[4][5][6]
Multitarget	Sorafenib	Drug Info	[4][5][6]
Cross References	3D Structure		
	Related Literature		
	On-Line Medical Dictionary		
Ref 1	Raf proteins and cancer: B-Raf is identified as a mutational target. Biochim Biophys Acta. 2003 Jun 5;1653(1):25-40. To Reference		

Figure 2. 5 Screenshots of detailed information page of ABL1 target.

The detailed drug information page contains drug name, synonyms, trade name, company, indications for use, 2D structure displayed, 3D structure in MOL format for download, InChI, InChIKey, canonical SMILES, therapeutic class, CAS number, Formula, Pubchem compound ID, Pubchem substance ID, ChEBI id, SuperDrug Anatomical Therapeutic Chemical (ATC) ID, SuperDrug CAS ID, therapeutic targets and references. Screenshots of drug information page taking Imatinib as an example are displayed in **Figure 2.6**. Cross-links to external chemical databases such as Pubchem, DrugBank, SuperDrug and ChEBI can be assessed through the IDs listed above and further information could be obtained.

TTD Drug ID: DAP000179			
Drug Information			
Name	Imatinib		
Synonyms	AC-524; 4-[(4-methylpiperazin-1-yl)methyl]-N-(4-methyl-3-[[4-(pyridin-3-yl)pyrimidin-2-yl]amino]phenyl)benzamide; Cgp 57148; 152459-95-5; ChEMBL941; sti-571; I01-1232; nchembio.282-comp6; alpha-(4-Methyl-1-piperazinyl)-3'-((4-(3-pyridyl)-2-pyrimidinyl)amino)-p-tolu-p-toluidide; Imatinib (INN); Glaxo; 4-(4-METHYL-PIPERAZIN-1-YLMETHYL)-N-[4-METHYL-3-(4-PYRIDIN-3-YL-PYRIMIDIN-2-YLAMINO)-PHENYL]-BENZAMIDE; Imatinib free base; BRD-K92723993-066-02-9; NCGC00159456-02; CID5291; MolPort-000-883-342; CCRIS 9076; benzamide, 4-[(4-methyl-1-piperazinyl)methyl]-N-[4-methyl-3-[[4-(3-pyridinyl)-2-pyrimidinyl]amino]phenyl]-; Imatinib; BIDD:GT0047; N-(3-(4-(pyridin-3-yl)pyrimidin-2-ylamino)-4-methylphenyl)-4-((4-methylpiperazin-1-yl)methyl)benzamide; AKOS000280662; STK617705; CGP 57148B; nchembio.117-comp23; Benzamide, 4-((4-methyl-1-piperazinyl)methyl)-N-(4-methyl-3-((4-(3-pyridinyl)-2-pyrimidinyl)amino)phenyl)-; AC1L1K0Z; Imatinib [INN:BAN]; EN002706; nchembio.162-comp5; 4-[(4-methylpiperazin-1-yl)methyl]-N-[4-methyl-3-[[4-(pyridin-3-yl)pyrimidin-2-yl]amino]phenyl]benzamide; LS-182208; DB00619; STI; FT-0083542; 4-[(4-methyl-1-piperazinyl)methyl]-N-[4-methyl-3-[[4-(3-pyridinyl)-2-pyrimidinyl]amino]phenyl]-benzamide methanesulfonate; Kinome_3724; NSC743414; 112GI019; NCGC00159456-04; 4-[(4-methylpiperazin-1-yl)methyl]-N-(4-methyl-3-[[4-(pyridin-3-yl)pyrimidin-2-yl]amino]phenyl)benzamide; 1iep; CHEBI:45783; Benzamide, 4-[(4-methyl-1-piperazinyl)methyl]-N-[4-methyl-3-[[4-(3-pyridinyl)-2-pyrimidinyl]amino]phenyl]- (9CI); nchembio.83-comp14; D08066; Imatinib Methanesulfonate; HMS2089D03; 1xbb; UNII-BKJ8M8G5HI; 4-[(4-Methyl-1-piperazinyl)methyl]-N-[4-methyl-3-[[4-(3-pyridinyl)-2-pyrimidinyl]amino]-phenyl]benzamide; DB03261; NCGC00159456-03; LS-187106; Glaxo (TN); STI571; STI 571		
Trade Name	Gleevec; Glivec		
Company	Novartis AG		
Indication	Chronic myelogenous leukemia [ICD9: 205.1 ICD10: C92.1]	Launched	[1]
	Glioma, lung, prostate, solid tumours [ICD9: 140-199, 191, 210-229 ICD10: C00-C75, C71, C7A, C7B, D10-D36, D3A]	Phase II	[1]
	Intestinal cancer & myeloid leukemia [ICD9: 152, 153, 205 ICD10: C17, C18, C92]	Phase III	[1]
Structure			

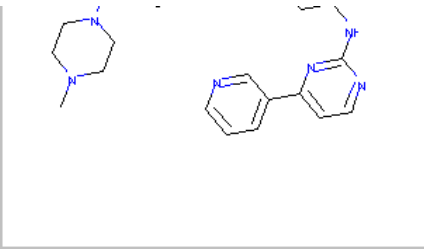
Structure				
	View the structure in Jmol			
	Click to save drug structure in 3D MOL format			
	Click to save drug structure in 2D MOL format			
InChI	1S/C29H31N7O/c1-21-5-10-25(18-27(21)34-29-31-13-11-26(33-29)24-4-3-12-30-19-24)32-28(37)23-8-6-22(7-9-23)20-36-16-14-35(2)15-17-36/h3-13,18-19H,14-17,20H2,1-2H3,(H,32,37)(H,31,33,34)			
InChIKey	KTUFNOKKBVMGRW-UHFFFAOYSA-N			
Canonical SMILES	CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5			
Therapeutic Class	Antineoplastic Agents			
CAS Number	CAS 152459-95-5			
Formular	C29H31N7O			
PubChem Compound ID	CID 5291.			
PubChem Substance ID	SID 584799.			
ChEBI	45783;			
SuperDrug ATC ID	L01XE01			
SuperDrug CAS ID	152459955;			
Target	Mast/stem cell growth factor receptor	Target Info	Inhibitor	[2]
	Mast/stem cell growth factor receptor	Target Info	Multitarget	[2]
	Platelet-derived growth factor receptor	Target Info	Inhibitor	[2]
	Platelet-derived growth factor receptor	Target Info	Multitarget	[2]
	Proto-oncogene tyrosine-protein kinase ABL1	Target Info	Inhibitor	[2]
	Proto-oncogene tyrosine-protein kinase ABL1	Target Info	Multitarget	[2]
Ref 1	Emerging treatments for pulmonary arterial hypertension. Expert Opin Emerg Drugs. 2006 Nov;11(4):609-19. To Reference			
Ref 2	A comparison of physicochemical property profiles of marketed oral drugs and orally bioavailable anti-cancer protein kinase inhibitors in clinical development. Curr Top Med Chem. 2007;7(14):1408-22. To Reference			

Figure 2. 6 Screenshots of detailed information page of drug Imatinib.

2.3.3 Biomarkers

In the main page of TTD database described above, other than the search by target and drug options, the most important search field is biomarker. And the incorporation of biomarker information is one of the most recent and essential

updates to TTD, in view of the current trends towards personalized drug treatment. The details of biomarker information in TTD will be highlighted and explained in detail here.

Overall 1,755 biomarkers for 365 disease conditions were collected, which included both process biomarkers (genetic mutations or alterations, gene amplification, and levels of proteins, gene expression, microRNAs, small molecules, or metabolites that capture a molecular/biochemical aspect of disease pathogenesis and the biological responses to the disease process and/or treatment) and global biomarkers (such as tumor sizes, brain structures in neurodegeneration, and shape of cells in anemia). These biomarkers may be searched in the “Search for biomarkers” field by using keywords or by selecting an ICD-9-CM/ICD-10-CM code (**Figure 2.1, Figure 2.3**).

Based on the literature descriptions, our collected biomarkers were classified into one or more of the following 11 classes: associative (disease correlation), antecedent (pre-illness risk identification), detective (disease early stage detection), classification (disease categorization and patient assignment for differential treatment), differentiative (differentiation of related diseases), diagnostic (recognition of overt diseases), monitoring (monitoring of disease state or treatment response), pharmacodynamic (examination of the biological basis of drug response

variations), prognostic (prediction of future disease course and response to therapy), surrogate (substitute of a clinical end point for predicting therapeutic benefit), and theragnostic (identification and monitoring of biochemical effects or mode of action of drug and downstream processes) classes.

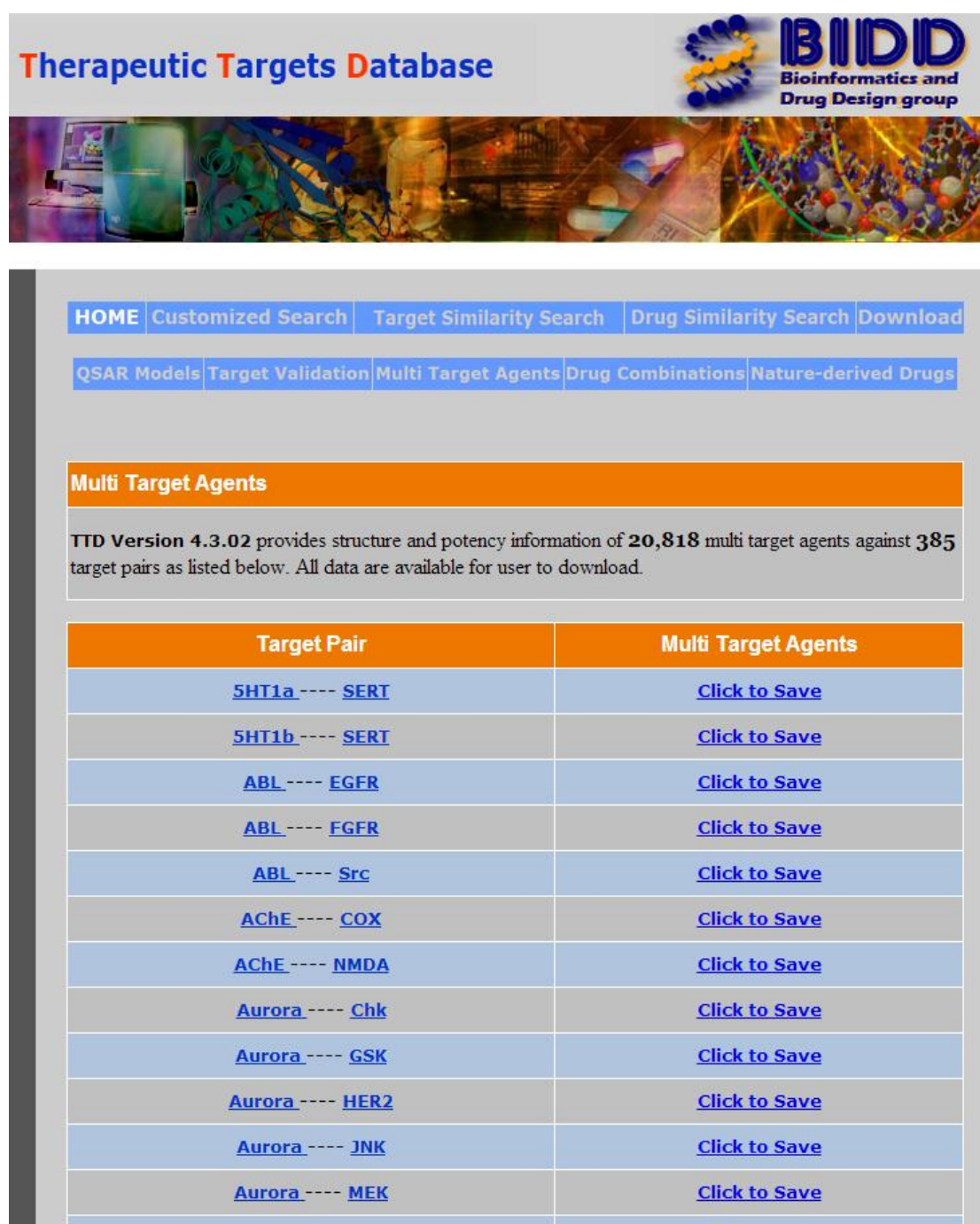
The biomarker detail information page contains the following fields whenever available, biomarker name, target ID (if this biomarker is a protein target in TTD), disease, ICD9 and ICD10 disease classification code, biomarker type (i.e. diagnostic, prognostic, etc) , molecule type (i.e. gene mutation, protein expression, microRNAs, small molecules, etc) , development phase (in clinical use or clinical trial), method to detect this biomarker (i.e. polymerase chain reaction (PCR), enzyme-linked immunosorbent assay (ELISA), mass spectrometry (MS), western blot, imaging, etc), measure (i.e. loss of function in a particular gene, elevated or reduced expression level of a biomarker, etc), specific use of the biomarker (i.e. to predict response to treatment, to predict survival rate, to monitor disease progression, to indicate adverse treatment effects, etc), conclusion of the biomarker study from literature, specific conclusion(usually give more details of the results of a biomarker study), treatment (if the biomarker is related to certain treatment or to predict the response to treatment), reference, uniprot ID and hyperlinks to gene bank, ChEMBL, Pfam, PDB, gene expression atlas, KEGG and gene ontology external databases. Screenshot of the biomarker information page can be seen in **Figure 2.6** using p53

in colon cancer as an example. A biomarker can have multiple functions in different disease conditions and, for each function, a table similar to the screenshot in **Figure 2.7** will be displayed if that biomarker is searched for.

2.3.4 Multi-target agents and drug combinations

The multi-target agents can be retrieved by clicking the ‘Multi-Target Agents’ field in the TTD home page, which leads to the TTD multi-target agents page where a user can download the multi-target agents against a specific target pair from the target pair list (**Figure 2.8**). For each target, if drugs against that particular target can also bind to other targets, then the multitarget agents are listed in the page and linked to the specific drug pages.

Similarly, the drug combinations can be retrieved by clicking on the “drug combination” tab. And drug-combination data, which include 72, 14 and 4 pharmacodynamically synergistic, additive, and antagonist combinations, and 19 and 7 pharmacokinetically potentiative and reductive combinations together with their mode of actions and combination mechanisms, are available for users to download.



Therapeutic Targets Database

BIDD
Bioinformatics and Drug Design group

HOME | Customized Search | Target Similarity Search | Drug Similarity Search | Download
QSAR Models | Target Validation | Multi Target Agents | Drug Combinations | Nature-derived Drugs

Multi Target Agents

TTD Version 4.3.02 provides structure and potency information of **20,818** multi target agents against **385** target pairs as listed below. All data are available for user to download.

Target Pair	Multi Target Agents
5HT1a ---- SERT	Click to Save
5HT1b ---- SERT	Click to Save
ABL ---- EGFR	Click to Save
ABL ---- FGFR	Click to Save
ABL ---- Src	Click to Save
AChE ---- COX	Click to Save
AChE ---- NMDA	Click to Save
Aurora ---- Chk	Click to Save
Aurora ---- GSK	Click to Save
Aurora ---- HER2	Click to Save
Aurora ---- JNK	Click to Save
Aurora ---- MEK	Click to Save

Figure 2.8 Screenshot of multi-target agents download page.

2.3.5 International Classification of Disease

From the TTD webpage (Figure 2.2), users can choose the “Search drugs and

targets by disease or ICD identifier ” field to search TTD target and drug entries by inputting disease names or selecting ICD-9-CM or ICD-10-CM codes. The TTD biomarker entries can be searched by ICD-9-CM or ICD-10-CM codes from the “Search for biomarkers” field. Users may also download from the TTD download page the lists of TTD target, drug and biomarker entries with the corresponding ICD-9-CM and ICD-10-CM codes.

ICD, short for International Classification of Disease, has been developed by WHO, sponsored by the United Nations, adopted by > 110 countries, and used by physicians, researchers, nurses, health workers, health information managers, policy-makers, insurers and health program managers for defining diseases, studying disease patterns, managing health care, monitoring outcomes and allocating resources (92, 93). ICD codes have been regularly revised to the current version of ICD-10 (92). But the previous version ICD-9 is still used by some organizations while proceeding with transition to ICD-10 (the expected completion date for the transition to ICD-10 in the United States is October 1st 2014) (95). ICD-10 is composed of 68,000 alphanumeric codes as compared to the 13,000 numeric codes of ICD-9, thus offering more comprehensive coverage and better representation of medical conditions (92). A number of nations have developed their own adaptations of the ICD codes. For instance, the United States have developed ICD-9 and ICD-10 clinical modification ICD-9-CM (17,000 codes) and ICD-10-CM (155,000 codes)

for covering additional morbidity details (96), which were used in TTD because of their more comprehensive coverage. **Table 2.2** provides the list of ICD-9-CM and ICD-10-CM code blocks together with the corresponding disease classes.

Table 2. 2 List of ICD-9-CM and ICD-10-CM code blocks and the corresponding classes of diseases and related health problems

ICD Code Chapter	ICD-9-CM Code Block	ICD-10-C M Code Block	Class of diseases or related health problems
I	<u>001–139</u>	A00-B99	Certain infectious and parasitic diseases
II	140–239	C00-D48	Neoplasms
III	279–289	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	<u>240–278</u>	E00-E90	Endocrine, nutritional and metabolic diseases
V	<u>290–319</u>	F01-F99	Mental and behavioral disorders
VI	<u>320–359</u>	G00-G99	Diseases of the nervous system
VII	<u>360–379</u>	H00-H59	Diseases of the eye and adnexa
VIII	<u>380–389</u>	H60-H95	Diseases of the ear and mastoid process
IX	<u>390–459</u>	I00-I99	Diseases of the circulatory system
X	<u>460–519</u>	J00-J99	Diseases of the respiratory system
XI	<u>520–579</u>	K00-K93	Diseases of the digestive system

XII	<u>680–709</u>	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	<u>710–739</u>	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	<u>580–629</u>	N00-N99	Diseases of the genitourinary system
XV	<u>630–679</u>	O00-O99	Pregnancy, childbirth and the puerperium
XVI	<u>760–779</u>	P00-P96	Certain conditions originating in the perinatal period
XVII	<u>740–759</u>	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	780–799	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	800–999	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	E000-E999	V01-Y98	External causes of morbidity and mortality
XXI	V01-V91	Z00-Z99	Factors influencing health status and contact with health services

The ICD-9-CM and ICD-10-CM codes were matched to the TTD target, drug and biomarker entries by the following procedures: First, automated word match was conducted for matching the disease name or names of each TTD target, drug or biomarker entry with the disease descriptions of each ICD codes. Secondly, each of the fully or partially matched TTD entry was manually checked to either validate the

match or to find the right ICD codes. Thirdly, manual search was conducted for those TTD entries without a single match. So far, ICD codes for 785 targets and 3,080 drugs that are linked to 732 disease conditions have been found.

The 1755 literature-reported biomarkers were also linked to ICD codes. By linking the ICD and biomarker codes, the relevant information may be conveniently accessed by clicking an icon beside a disease name/code in an E-health system. Most biomarkers are molecular-based, which present an educational challenge for those users who are only familiar with the histopathology-based disease-classification systems. Linking ICD and biomarker codes enables easy cross-links to bioinformatics resources for genomic, structural, pathway and functional information.

In addition, the linking of ICD codes to targeted therapeutics information like biomarkers makes it possible to analyze the disease coverage pattern of such therapeutics. And chapter 5.2 gives an example of such analysis on non-invasive molecular biomarkers in detail.

A new ICD version ICD-11 is in development and scheduled for endorsement by WHO in 2015 (97), which offers more refined disease classification based on more recent scientific understanding of the disease mechanisms. For instance, small cell

lung cancer, which represents approximately 13% of all lung cancer diagnoses (98), is not explicitly classified in the ICD-10 and earlier ICD versions but is now explicitly represented in the ICD-11 beta draft. Therefore, ICD-11 is expected to be more useful for developing a more refined disease classification system for stratified and personalized medicine. Effort will be made to upgrade TTD to the ICD-11 version upon its official release. Moreover, suggestions to be made to ICD codes for more refined disease classification will be discussed in chapter 5.1.

2.4 Future work

The efforts in the discovery and application of targeted therapeutics increasingly involve collective efforts from multiple bench-to-clinic communities (24-26) and are moving more and more towards stratified and personalized medicines (27-29). The drug, target, biomarker, and other relevant chemical, biological, pharmaceutical and clinical data need to be more integrated and be made easily accessible by the multiple bench-to-clinic communities.

The updated TTD database continues to be a primary source providing integrated information about targeted therapeutics with easily accessible features by biological researchers, pharmaceutical industry and medical practitioners. Over 130,000 clicks to TTD database since its construction and over 44000 click counts of visit to TTD

since its last update demonstrate its popularity and important role in the field of pharmaceutical research.

Continuous efforts will be made to expand the linkage of the ICD and ATC codes to drugs, efficacy targets and biomarkers and to provide the latest and comprehensive information about the drugs, efficacy targets and biomarkers for better serving the multiple bench-to-clinic communities in their collective efforts for the discovery, investigation, application, monitoring and management of targeted therapeutics.

Potential biomarkers, particularly multi-markers, have also been predicted from the genetic and gene expression data of patients by using such computational methods as the principal components analysis feature selection method (99), weighted voting classification feature selection method (100), hierarchical clustering feature selection method (101), differentially expressed genes method (102, 103), and machine learning feature selection methods (104, 105). These potential biomarkers may also be included in TTD and other biomarker databases for facilitating their future exploration.

Chapter 3: Methods to learn from known drugs and inhibitors for the design of multi-target small molecule drugs

Information of known drugs and inhibitors is a valuable source to facilitate the design of traditional multi-target small molecule drugs. Machine learning methods can be implemented as virtual screening tools to learn from physiochemical properties of the known inhibitors and make predictions to select virtual hits that act on multiple targets simultaneously. The performance of the machine learning methods for prediction of multi-target agents was evaluated in Chapter 3.1. Through the hierarchical clustering method described in Chapter 3.2, drugs, inhibitors and similar compounds can be clustered and possible drug distribution patterns could be learnt by analysing the cluster patterns.

3.1 Evaluation of Hit and Target Selection Performance of Machine Learning Multi-Target Virtual Screening Methods

In this work, the multi-target hit and target selection performance of 3 extensively-used machine learning methods, Combi-SVM, combinatorial k-Nearest Neighbour (Combi-kNN) and combinatorial Probabilistic Neural Network (combi-PNN) in the VS of dual inhibitors of 29 target pairs of high, intermediate and low similarity levels between their drug-binding domains was systematically

evaluated. These target pairs cover 8 therapeutically explored biochemical classes including kinases, proteases, transporters, GPCRs, reductases, synthases, cytokines, and DNA-binding proteins. The yield of multi-target hits of the three VS methods was rigorously tested by using individual target inhibitors as training datasets (all known dual inhibitors are excluded) and the known dual inhibitors as independent testing datasets of these VS methods. Such tests are particularly useful for testing the capability of these methods in identifying multi-target inhibitors without explicit knowledge of multi-target agents (23). Target selectivity of these VS methods were assessed by measuring the false hit rates in misclassifying individual-target inhibitors of a target pair and inhibitors of the other targets in the same biochemical class as dual inhibitors of the target pair. Moreover, the ability of these VS methods in searching large compound libraries were evaluated by using them to screen 17 million and 168,000 compounds from the PubChem and MDL Drug Data Report (MDDR) databases with particular focus on estimating false-hit rates in screening these libraries.

3.1.1 Method

3.1.1.1 Datasets and molecular descriptors

Based on sequence similarity between their drug binding domains obtained from PFAM, the 29 target pairs were classified into 3 classes, namely low, intermediate and high similarity classes. According to Rost's finding that proteins with >40%

sequence identity unambiguously distinguish similar and non-similar structures while the distinction signal gets blurred in the twilight zone of 20-30% (106) , the target pairs were classified into high, intermediate and low similarity classes with their drug-binding domains at sequence identity levels of >40%, 20-40% and <20% respectively (**Table 3.1**). The high-similarity target-pairs include SERT-NET, Src-Lck, VEGFR2-Lck, CDK1-CDK2 and PDGFR-FGFR. The intermediate-similarity target-pairs include EGFR-Src, CDK2-GSK3, MMP2-MMP3, EGFR-FGFR, CDK1-GSK3, EGFR-PDGFR, Aurora-GSK3, PDGFR-Src, Aurora-Met, Aurora-HER2 and CDK1-VEGFR2. The low-similarity target-pairs include PKC-Topoisomerase, SERT-5HT1b, AggreCANase-MMP1, DHFR-Thymidylate Synthase, AggreCANase-MMP9, AggreCANase-TNF α , AggreCANase-MMP2, SERT-5HT1a, HER2-MMP2, HER2-MMP9, and SERT-H3.

Table 3. 1 Datasets of individual-target and multi-target inhibitors of the target-pairs used for developing and testing machine learning multi-target inhibitor virtual screening tools. Additional sets of 17 million PubChem compounds and 168,000 MDDR active compounds were also used for the test.

Target Pair			Inhibitors in Training Sets		Inhibitors in Testing Sets	
Target A – Target B	Drug-binding domain similarity level (sequence identity)	Biochemical class	No of inhibitors of A that are non-inhibitor of B	No of inhibitors of B that are non-inhibitor of A	No of multi-target agents of A and B	No of inhibitors of other targets in the same biochemical class
SERT-NET	High-similarity (72.3%)	Transporter-Transporter	1125	1410	101	
Src-Lck	High-similarity (67.1%)	Kinase-Kinase	804	450	56	4906
VEGFR2-Lck	High-similarity (66.9%)	Kinase-Kinase	1232	445	61	4515
CDK1-CDK2	High-similarity (66.8%)	Kinase-Kinase	484	650	174	4945
PDGFR-FGFR	High-similarity (40.4%)	Kinase-Kinase	450	233	230	5339
EGFR-Src	Intermediate-similarity (37.4%)	Kinase-kinase	1262	748	112	4083
CDK2-GSK3	Intermediate-similarity (36.8%)	Kinase-Kinase	749	722	75	4704
MMP2-MMP3	Intermediate-similarity	Protease-Protease	674	662	12	1918

	(35.5%)					
EGFR-FGFR	Intermediate-similarity (33.1%)	Kinase-Kinase	1303	392	71	4486
CDK1-GSK3	Intermediate-similarity (33.1%)	Kinase-Kinase	503	642	155	4955
EGFR-PDGFR	Intermediate-similarity (28.0%)	Kinase-Kinase				
Aurora-GSK3	Intermediate-similarity (29.6%)	Kinase-Kinase	672	1192	44	3147
PDGFR-Src	Intermediate-similarity (21.3%)	Kinase-Kinase	492	672	188	4844
Aurora-Met	Intermediate-similarity (23.7%)	Kinase-Kinase	698	442	18	3834
Aurora-HER2	Intermediate-similarity (20.6%)	Kinase-Kinase	690	937	26	3331
CDK1-VEGFR2	Intermediate-similarity (21.6%)	Kinase-Kinase	651	1285	41	4312
PKC-Topoisomerase	Low-similarity (15.5%)	Kinase-DNA binding protein	1156	805	9	
SERT-5HT1b	Low-similarity (15.1%)	Transporter-GPCR	1894	917	57	
Aggrecanase-MMP1	Low-similarity (11.9%)	Protease-Protease	252	1289	44	1692
DHFR-thymidylate synthase	Low-similarity (10.6%)	Reductase-Synthase	1465	557	139	

Aggrecanase-MMP9	Low-similarity (10.1%)	Protease-Protease	279	340	17	2668
Aggrecanase-TNFalpha	Low-similarity (9.0%)	Protease-Cytokine	281	68	15	3008
Aggrecanase-MMP2	Low-similarity (10.7%)	Protease-Protease	286	676	10	2344
SERT-5HT1a	Low-similarity (8.0%)	Transporter-GPCR	1679	1144	216	
HER2-MMP2	Low-similarity (4.8%)	Kinase-Protease	936	659	27	6564
HER2-MMP9	Low-similarity (2.4%)	Kinase-Protease	951	345	12	6895
SERT-H3	Low-similarity (1.7%)	Transporter-GPCR	1804	1689	147	

Individual-target and dual-target inhibitors for the 29 target pairs (**Table 3.1**), each with $IC_{50} \leq 20 \mu M$, were collected from the literatures and public databases such as ChEMBL (107) and BindingDB (108) databases. As few non-inhibitors have been reported, putative non-inhibitors of each target were generated by using our method reported in our earlier publications (109, 110). In our method, 13.56 million PubChem and 168 thousand MDDR compounds were clustered into 8,993 compound families(111). The number of our derived compound families are consistent with the reported 12,800 groups of topologically close structures for 26.4 million compounds of up to 11 atoms (112), and 2,851 clusters for 171,045 natural products (113). 5 representative compounds were selected from each family that contain no known individual-target and dual-target inhibitors as the putative non-inhibitors for developing the three VS tools. This approach has the risk of wrong inclusion of the compound families that contain undiscovered multi-target and individual-target inhibitors into the non-inhibitor training dataset. The maximum possible “wrong” classification rate arising from these mistakes has been estimated at <13% even in the extreme and unlikely cases that all of the undiscovered single-target and multi-target agents as well as the known multi-target agents are misplaced into the non-inhibitor class (110, 114, 115). The noise level generated by up to 13% “wrong” negative compound family representation is expected to be substantially smaller than the noise level tolerated by machine learning methods such as SVM (116).

Each compound (including putative non-inhibitors) was represented by 98 1D and 2D molecular descriptors derived from our own software (117), which are composed of 18 descriptors in the class of simple molecular properties, 3 descriptors in the class of chemical properties, 35 descriptors in the class of molecular connectivity and shape, 42 descriptors in the class of electro-topological state. These descriptors have been extensively used in deriving structure-activity relationships (118), quantitative structure activity relationships (119) and machine learning VS methods for individual-target (110, 114, 115) and multi-target (23) agents.

3.1.1.2 Support vector machines

Based on the structural risk minimization principle of statistical learning theory (120), SVM performs well consistently with good classification capability, fast classification speed, low over-fitting risk, relative insensitivity to sample redundancy and ability to work on structurally diverse large datasets (121, 122). Though the performance of SVM is limited by the insufficient knowledge of known inhibitors for many targets, it is a useful tool to complement other virtual screening tools with comparable performances or improvement in aspects like reduced false-hit rates.

A linear SVM model tries to construct a hyper-plane that perfectly separates the active and inactive classes of compounds, represented by vectors of their molecular

descriptors in the multidimensional feature space, and maximizes the margin, defined as the closest distance from any vector point to the hyper-plane.

Mathematically, the hyper-plane satisfies the following conditions:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \text{ Active class}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \text{ Inactive class}$$

where y_i is the class index, \mathbf{w} is a vector normal to the hyper-plane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyper-plane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of \mathbf{w} . Based on \mathbf{w} and b , a given vector \mathbf{x} can be classified by $f(\mathbf{x}) = \text{sign}[(\mathbf{w} \cdot \mathbf{x}) + b]$. A positive or negative $f(\mathbf{x})$ value indicates that the vector \mathbf{x} belongs to the active or inactive class respectively

For the classification of compounds with diverse structures, a nonlinear SVM is frequently used, as the input vectors are not linearly separable. Kernel functions are used to map input vectors into a higher dimensional feature space that can be linearly separated. We used the radial basis function kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2}$, which are commonly used and proven to have consistent better performance over other kernel function (110, 114, 115). Linear SVM was then applied to this feature space based on the following decision function

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^l \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b\right), \text{ where the coefficients } \alpha_i^0 \text{ and } b \text{ were determined}$$

by maximizing the following Lagrangian expression:

$$\sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{under the conditions} \quad \alpha_i \geq 0 \quad \text{and}$$

$$\sum_{i=1}^l \alpha_i y_i = 0. \text{ A positive or negative } f(\mathbf{x}) \text{ value indicates that the vector } \mathbf{x} \text{ belongs to}$$

the active or inactive class respectively. Our SVM VS models were developed by using a hard margin $c=100,000$ and their σ values are in the range of 0.1-2. In terms of the numbers of true positives TP (true inhibitors), true negatives TN (true non-inhibitors), false positives FP (false inhibitors), and false negatives FN (false non-inhibitors), the yield and false-hit rate are calculated by $TP / (TP+FN)$ and $FP / (TP+FP)$ respectively.

3.1.1.3 k-Nearest Neighbour

k-nearest neighbour (k-NN) is a classification method based on the nearest input training vectors in the multidimensional feature space. It measures the distance of a to-be-classified vector \mathbf{x} and each individual vector \mathbf{x}_i in the training set, and the unknown vector is assigned to the class which majority of its k nearest neighbours belong to. (123, 124) The most common distance metric used is Euclidean distance,

calculated using the formula: $D = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2}$. k vectors nearest to the vector \mathbf{x} are

used to determine its class $\hat{f}(\mathbf{x}) \leftarrow \arg \max_{v \in V} \sum_{i=1}^k \delta(v, f(\mathbf{x}_i))$, where

$\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ if $a \neq b$, $\arg \max$ is the maximum of the function,

V is a finite set of vectors $\{v_1, \dots, v_s\}$ and $\hat{f}(\mathbf{x})$ is an estimate of $f(\mathbf{x})$ and is assigned to the same class as the most frequent class of the k nearest neighbours.

The best parameter k of constructed k -NN models is chosen to be in the range of $k=1$ or 3 or 5 or 9 , based on the highest yield of dual inhibitor prediction.

3.1.1.4 Probabilistic Neural Network

Probabilistic Neural Network (PNN) is a classification method based on Bayes' optimal decision rule (125): $h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x})$, where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification and $f_i(x)$ and $f_j(x)$ are the probability density function for class i and j respectively.

An unknown vector \mathbf{x} is classified into population i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator,

$$g(\mathbf{x}) = \frac{1}{n\sigma} \sum_{i=1}^n W\left(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma}\right)$$

where n is the sample size, σ is a scaling parameter which defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weight function which has its largest value at $d = 0$ and $(\mathbf{x} - \mathbf{x}_i)$ is the distance between the unknown vector and a vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos for the multivariate case.

$$g(x_1, \dots, x_p) = \frac{1}{n\sigma_1 \dots \sigma_p} \sum_{i=1}^n W\left(\frac{x_1 - x_{1,i}}{\sigma_1}, \dots, \frac{x_p - x_{p,i}}{\sigma_p}\right)$$

The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated and satisfies the conditions required by Parzen's estimator.

Thus the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \exp\left(-\sum_{j=1}^p \left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2\right)$$

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. There are 4 layers in a PNN. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparameteric estimator. The summation layer has a neuron for each class and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. The single neuron in the output layer then estimates the class of the unknown vector \mathbf{x} by comparing all the probability density function from the summation neurons and choosing the class with the highest probability density function. The parameters of the developed PNN models for the evaluated targets are in the range of $\delta=0.001\sim0.015$. And the optimal parameter is chosen based on the highest yield of dual inhibitor prediction. Only two classes (active or inactive) are

trained in the PNN models to predict inhibitors.

3.1.2 Results and discussion

3.1.2.1 Dual inhibitor yields

The VS performances of the three VS methods for high, intermediate and low similarity target pairs are shown in **Table 3.2** respectively and in **Figure 3.1**. All three VS methods showed comparable dual-inhibitor yields for the target pairs at all similarity levels. Specifically, the dual inhibitor yields of Combi-SVM, Combi-kNN and Combi-PNN are in the range of 17.65%-77.80%, 10.90%-88.89%, and 38.1-100% in searching low similarity target pairs, 14.63%-73.10%, 5.56%-66.7%, and 14.63%-75.00% in searching intermediate similarity target pairs, and 38.26%-75.00%, 16.09%-67.86%, and 21.30%-83.93% in searching high similarity target pairs respectively. These yields are comparable to the yields of dual kinase inhibitors and dual-target serotonin reuptake inhibitors produced by Combi-SVM in our earlier studies (23). These yields are also comparable to that of QSAR method reported in the literature. A recently developed multi-target kinase inhibitor QSAR model correctly identified the dual targets of 2 (66.6%) of the 3 dual kinase inhibitors (EGFR-Lck inhibitor Pelitinib, and VEGFR2-PDGFR inhibitors Sunitinib and Sorafenib) tested among several kinase inhibitors (47).

Table 3. 2 Virtual screening performance of combinatorial SVMs for identifying dual-target inhibitors of high similarity target pairs

Target Pair	Method	Virtual Screening Performance
-------------	--------	-------------------------------

Target A – Target B		Multi-target inhibitors	Inhibitors of individual target of the target-pair inactive against the other target of the target-pair		Inhibitors of other targets of the same biochemical classes of the target-pair	All 168,000 MDDR compounds	17 million PubChem compounds
		Yield	False hit rate for inhibitors of target A	False hit rate for inhibitors of target B	False hit rate	Virtual hit rate	Virtual hit rate
SERT-NET	Combi-SVM	49.5%	22.4%	29.8%	2.4%	0.12%	0.035%
	kNN	59.40%	19.80%	25.10%		0.580%	
	PNN	57.40%	52.30%	38.40%		3.140%	
Src-Lck	Combi-SVM	75.00%	16.42%	7.61%	0.84%	0.034%	0.011%
	kNN	67.86%	14.55%	22.02%		0.140%	
	PNN	83.93%	23.26%	36.21%		0.465%	
VEGFR2-Lck	Combi-SVM	52.46%	29.21%	6.49%	3.39%	0.104%	0.036%
	kNN	42.62%	9.90%	30.79%		0.208%	
	PNN	54.10%	20.70%	38.88%		0.865%	
CDK1-CDK2	Combi-SVM	52.3%	39.2%	48.1%	3.4%	0.075%	0.022%
	kNN	16.09%	21.83%	18.71%		0.170%	
	PNN	21.30%	34.51%	32.52%		0.577%	
PDGFR-FGFR	Combi-SVM	38.26%	13.78%	22.75%	4.44%	0.056%	0.013%
	kNN	36.96%	13.11%	30.47%		0.139%	
	PNN	60.87%	28.00%	48.50%		0.665%	
EGFR-Src	Combi-SVM	26.8%	12.9%	11.1%	1.49%	0.096%	0.033%
	kNN	51.79%	12.38%	25.63%		0.243%	
	PNN	67.86%	27.27%	42.32%		0.968%	
CDK2-GSK3	Combi-SVM	34.67%	8.00%	9.35%	0.77%	0.071%	0.016%
	kNN	30.67%	15.58%	15.65%		0.164%	

	PNN	52.00%	24.10%	29.22%		0.571%	
MMP2-MMP3	Combi-SV M	66.67%	27.45%	23.41%		0.13%	0.018%
	kNN	66.67%	36.65%	35.95%		0.714%	
	PNN	75.00%	49.55%	33.38%		0.738%	
EGFR-FGFR	Combi-SV M	40.85%	7.37%	8.16%	1.38%	0.071%	0.015%
	kNN	52.50%	12.36%	22.45%		0.169%	
	PNN	74.65%	22.49%	43.62%		0.783%	
CDK1-GSK3	Combi-SV M	30.32%	8.00%	9.35%	1.15%	0.037%	0.018%
	kNN	25.81%	12.40%	10.59%		0.082%	
	PNN	45.81%	31.20%	27.26%		0.566%	
EGFR-PDGFR	Combi-SV M	27.60%	9.20%	14.30%	1.88%	0.100%	0.031%
	kNN	34.50%	14.74%	27.17%		0.274%	
	PNN	51.72%	21.35%	45.98%		0.662%	
Aurora-GSK3	Combi-SV M	47.73%	13.24%	4.87%	0.13%	0.118%	0.053%
	kNN	36.36%	9.38%	9.40%		0.152%	
	PNN	56.82%	31.40%	19.55%		0.118%	
PDGFR-Src	Combi-SV M	38.3%	25.8%	11.6%	1.81%	0.10%	0.021%
	kNN	40.40%	34.96%	19.23%		0.242%	
	PNN	72.34%	46.14%	40.80%		0.768%	
Aurora-Met	Combi-SV M	16.7%	3.6%	9.3%	0.8%	0.018%	0.0095%
	kNN	5.56%	6.45%	6.79%		0.052%	
	PNN	22.22%	9.74%	18.10%		0.217%	
Aurora-HER2	Combi-SV M	73.1%	20.4%	13.7%	1.1%	0.1%	0.034%
	kNN	34.62%	19.86%	15.90%		0.167%	
	PNN	61.54%	33.19%	28.18%		0.550%	
CDK1-VEGFR2	Combi-SV M	14.63%	0.78%	1.73%	4.77%		
	kNN	14.63%	15.09%	10.76%		0.200%	
	PNN	14.63%	1.24%	0.68%		0.840%	
PKC-Topoisomerase	Combi-SV M	77.8%	3.2%	0.35%		0.022%	0.0065%
	kNN	88.89%	3.98%	5.96%		0.085%	

	PNN	66.67%	3.72%	6.34%		0.109%	
SERT-5HT1b	Combi-SV M	38.6%	13.8%	37.9%	3.4%	0.24%	0.035%
	kNN	31.60%	14.10%	32.60%		0.750%	
	PNN	45.60%	4.70%	30.30%		2.830%	
Aggrecanase-MMP1	Combi-SV M	50.00%	16.67%	5.35%	3.31%	0.065%	0.008%
	kNN	31.82%	23.41%	5.35%		0.139%	
	PNN	61.36%	43.25%	15.13%		0.978%	
DHFR-thymidylate synthase	Combi-SV M	39.6%	27.5%	33.4%		0.14%	0.02%
	kNN	33.81%	21.37%	24.42%		0.179%	
	PNN	46.76%	32.70%	33.75%		0.323%	
Aggrecanase-MMP9	Combi-SV M	17.65%	5.38%	4.41%	2.25%	0.030%	0.004%
	kNN	11.76%	6.45%	6.76%		0.062%	
	PNN	47.06%	33.69%	11.47%		0.370%	
Aggrecanase-TNFalph a	Combi-SV M	46.67%	2.14%	1.47%	0.70%	0.012%	0.000%
	kNN	33.33%	2.14%	5.88%		0.023%	
	PNN	100.00%	4.98%	4.41%		0.011%	
Aggrecanase-MMP2	Combi-SV M	60.00%	7.34%	10.50%	1.83%	0.027%	0.004%
	kNN	50.00%	7.69%	9.47%		0.077%	
	PNN	80.00%	25.87%	14.50%		0.377%	
SERT-5HT1a	Combi-SV M	47.7%	15.4%	19.4%	7.1%	0.28%	0.054%
	kNN	34.30%	16.60%	24.30%		0.830%	
	PNN	45.40%	38.90%	34.30%		3.400%	
HER2-MMP2	Combi-SV M	74.07%	2.03%	1.97%	0.49%	0.010%	0.006%
	kNN	29.63%	1.60%	2.28%		0.032%	
	PNN	77.78%	4.81%	3.79%		0.175%	
HER2-MMP9	Combi-SV M	41.67%	2.31%	2.03%	0.45%	0.001%	0.001%
	kNN	50.00%	1.05%	2.61%		0.007%	
	PNN	75.00%	3.05%	2.61%		0.035%	
SERT-H3	Combi-SV M	25.9%	5.4%	8.2%	3.5%	0.067%	0.028%
	kNN	10.90%	9.00%	8.50%		0.410%	
	PNN	38.10%	25.50%	22.20%		2.350%	

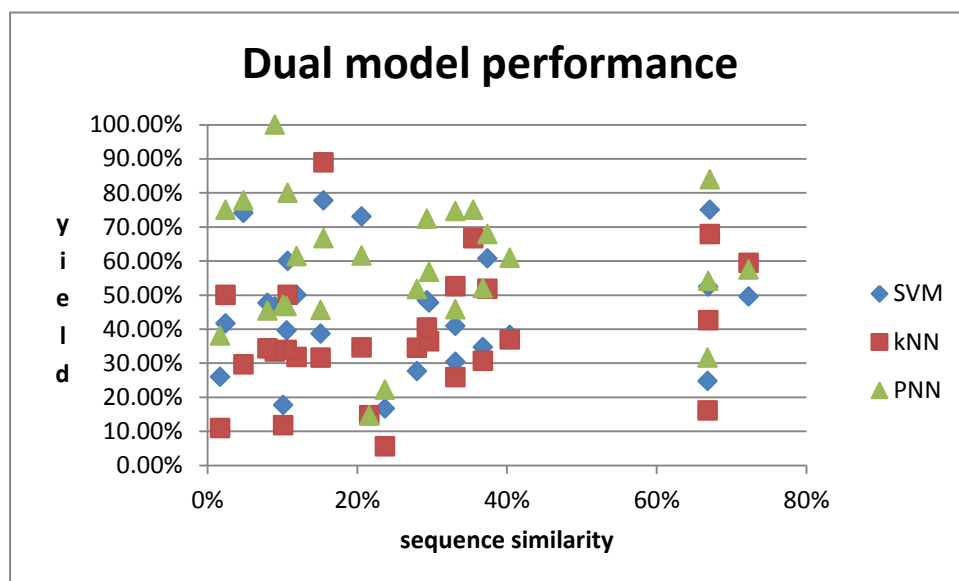


Figure 3. 1 Dual model performance of three machine learning methods.

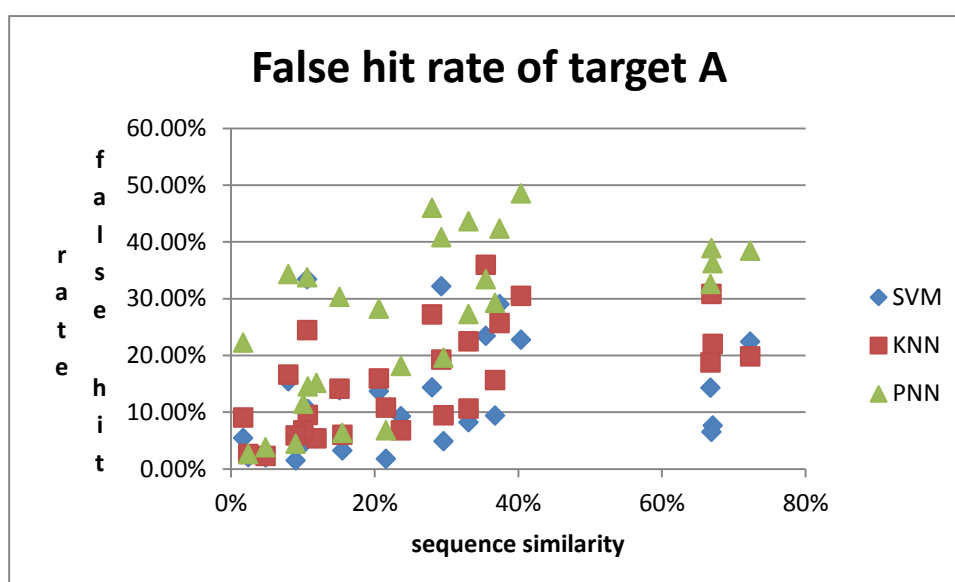
As shown in **Figure 3.1**, the dual inhibitor yields tend to show larger variations at decreasing similarity between the drug-binding domains of the target pairs. This suggests that it is more difficult to produce consistent dual inhibitor yields for lower similarity target pairs. Cases of disagreement between sequence-based similarity and binding site similarity have been reported (126). In particular, some protein pairs with very low similarity at the sequence level may have high levels of similarity in their binding site surface characteristics (127). On the other hand, it has been found that some multi-target agents bind to targets of different families with different binding site structures by adopting substantially different conformations (induced fit) and relying on additional help such as metal binding (128-130). These factors may not be fully captured by currently available molecular descriptors and machine learning methods. Therefore, dual inhibitors of low similarity target pairs with

higher binding site similarity are expected to be more easily identified by machine learning methods than those of the target pairs with lower binding site similarity, which may partly contribute to the larger variations of dual inhibitor yields for low similarity pairs.

3.1.2.2 Target selectivity

Target selectivity against individual target inhibitors of the same target pair was tested by using the three machine learning methods to screen the 68-1894 individual target inhibitors of each target-pair to determine the percentage of individual target inhibitors of the same target pair incorrectly predicted as dual inhibitor of the target pair. As shown in **Table 3.2**, Combi-SVM, Combi-kNN and Combi-PNN misidentified 0.35%-37.90%, 1.05%-32.60%, and 2.61-43.25% of the individual-target inhibitors as dual-inhibitors for low similarity pairs, 0.78%-25.80%, 6.45%-36.65%, and 0.68-49.55% for intermediate similarity pairs, and 6.49%-48.10%, 9.90%-30.79%, and 20.70-52.30% for high similarity pairs, respectively. Therefore, all three methods are reasonably selective in distinguishing dual inhibitors from individual-target inhibitors of the same target pairs. As shown in **Figure 3.2**, the selectivity of all three methods against individual-target inhibitors tends to be significantly decreased when similarity level of the target pairs is increased. This is consistent with the findings from several reported target selectivity studies. It has been reported that inhibitors tend to become less selective to binding

sites with less distinct physicochemical properties (131). The structure-activity landscapes of the bioactive compounds of closely related targets are expected to include overlapping and distinct regions of multi-target agents many of which with structures similar to the individual-target inhibitors (132). These factors may make it harder for machine learning methods to distinguish dual-target and individual-target inhibitors for high similarity target pairs. Two additional factors may contribute to the misclassification of individual target inhibitors as dual inhibitors. First, the three methods were trained by using individual target inhibitors only, which are not expected to fully distinguish dual inhibitors from individual-target inhibitors. Secondly, some of the misidentified individual target inhibitors may be true dual inhibitors not yet experimentally tested for multi-target activities.



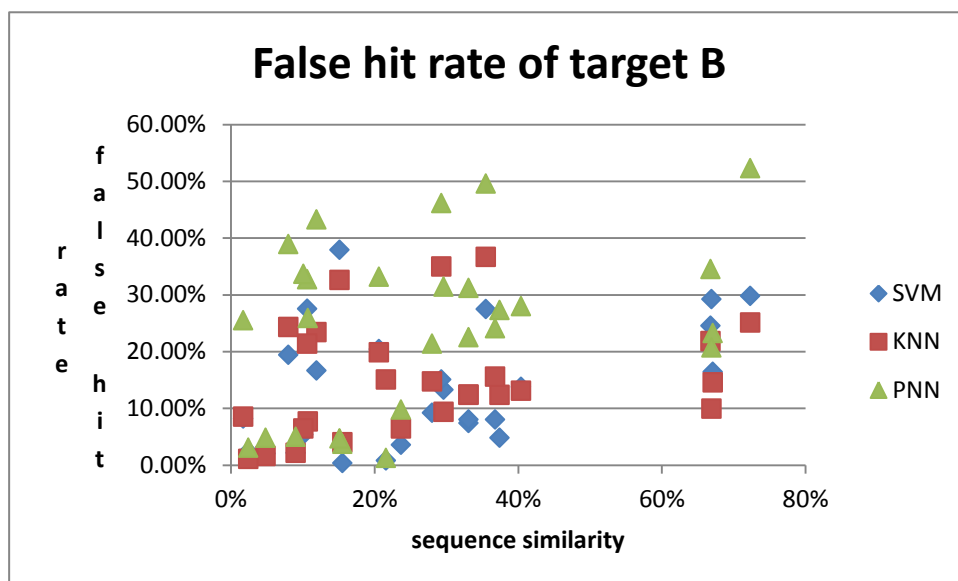


Figure 3. 2 Selectivity of three methods against individual-target inhibitors

Target selectivity was further tested by using SVM method to screen the inhibitors of the other targets in the same biochemical class studied in this project outside the target pair (**Table 3.2**) We found that small percentages of 0.45%-7.10% of the individual-target inhibitors were misidentified by Combi-SVM as dual-inhibitors for low similarity pairs, 0.13%-4.77% for intermediate similarity pairs, and 0.84%-4.44% for high similarity pairs respectively. Compared to their selectivity against individual target inhibitors, Combi-SVM is significantly more selective against inhibitors of other targets in the same biochemical class outside the target-pair, and the selectivity is insensitive to the level of similarity of the target pairs. This is consistent with the conclusions from extensive studies of kinase selectivity profiles of kinase inhibitors. Screening of two scaffold groups of 118 compounds against a panel of 353 kinases has shown that each scaffold has distinct kinase selectivity profile with selective inhibitory activity against a small number of kinases (133).

Global kinase target profiling of several BCR-ABL kinase inhibitors imatinib, nilotinib, dasatinib, bosutinib, and INNO-406 has shown that each of these inhibitors exhibits overlapping but distinct inhibition profiles across the whole kinase panel (134-136). Although kinase inhibitors have a propensity to cross-interact with multiple kinases, not all kinases are equally likely to interact with small molecules (137). An earlier analysis of corporate data suggests that kinase frequent hitters are far fewer in numbers than kinase selective inhibitors (138). These studies have consistently shown that kinase inhibitors have no apparent propensity to cross-interact with other kinases of similar drug-binding domain sequences. One may further speculate that inhibitors of other biochemical classes behave in a similar way.

3.1.2.3 Virtual screening performance in searching MDDR database

As shown in **Table 3.2**, the numbers of dual inhibitor virtual-hits identified by Combi-SVM, Combi-kNN and Combi-PNN and the corresponding virtual-hit rates in screening 168,000 MDDR compounds are 0.00%-0.28%, 0.01%-0.83% and 0.01%-3.40% for low similarity pairs, 0.02%-0.12%, 0.05%-0.71%, and 0.12%-0.97% for intermediate similarity pairs, and 0.03%-0.12%, 0.14%-0.58%, and 0.47%-3.14% for high similarity pairs respectively. As shown in **Figure 3.3**, the virtual hit rates of the 3 methods are relatively insensitive to the similarity levels of the target pairs. One possible reason for the low sensitivity to target pair similarity is that most of the MDDR compounds are significantly different in structural and physicochemical

properties to the dual inhibitors and individual target inhibitors of the evaluated target pairs.

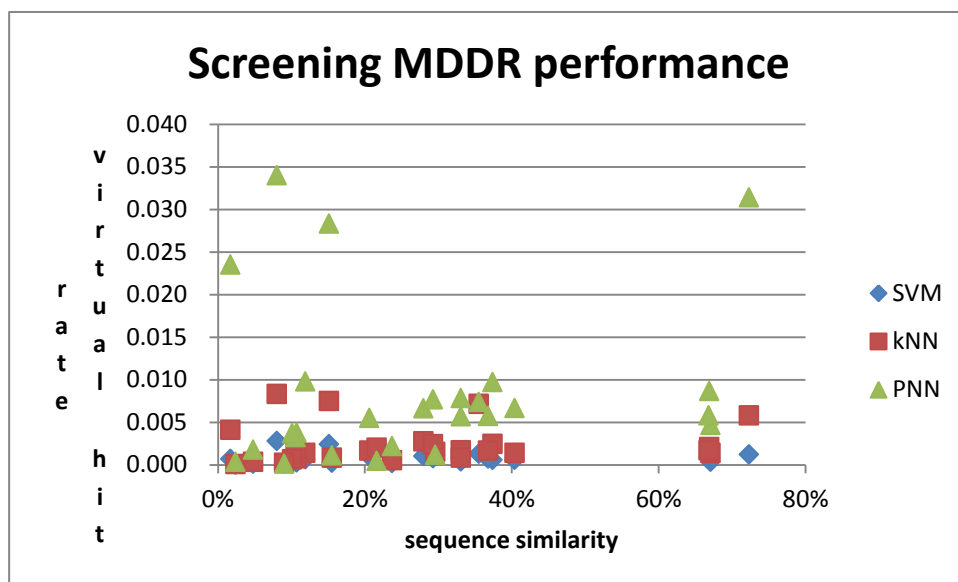


Figure 3. 3 The virtual hit rates of three machine learning methods to screen MDDR

Given the possibility that some of the identified MDDR virtual hits may be true dual inhibitors, the true false hits rates of the 3 methods are likely smaller than the computed virtual hit rates. Therefore, the false hit rates of Combi-SVM, Combi-kNN and Combi-PNN can be estimated as $\leq 0.00\%-0.28\%$, $\leq 0.01\%-0.83\%$ and $\leq 0.01\%-3.4\%$ in screening MDDR compounds respectively. These rates are comparable and in some cases better than the false-hit rates of 0.02%-0.37% and 0.05%-0.35% produced by some of the machine learning methods and molecular docking methods reported in the literature (23).

3.1.3 Future work

As discussed in previous section, the virtual screening performance of machine learning approaches needs further improvement. Among the many aspects that could be improved, a lower false hit rate is desired when screening large databases like Pubchem. This is to ensure that a sufficient percentage of virtual hits are true hits so as to reduce the costs of wet lab validations.

To decrease the false hit rate, a better method to generate putative non-inhibitor is necessary. The Pubchem database has been updated, so by keeping our SVM-formatted Pubchem screening library up to date, the putative non-inhibitors generated could be more representative of the whole chemical space.

In the above method section, 13.56 million Pubchem compounds were clustered to over 8000 families in order to generate the putative negatives. Due to the significant increase in the size of Pubchem compound structures, efforts have been made to recluster the updated Pubchem. 29.7 million Pubchem compounds with distinct structures were cleaned, formatted and clustered. The more than two fold increase in the number of Pubchem compounds was beyond the computational power of our current computers using k-means clustering. After running consecutively for 2 months, the k-means clustering method still did not generate any clustering results. Hence, in house script was written in Fortran to split the Pubchem chemical space into parts and then k-means was applied to each part to divide the part of chemical

space into smaller clusters of compounds families. The idea of this Fortran script was to find the center of the chemical space first and then calculate the Euclidian distance between each data points to the center. Groups of points within a certain range of distance to the center points, which can be visualized as a spherical ring if in 3D space, will be classified as one part of chemical space. The 29.7 million Pubchem compound space was split into 168 such parts. And then each part was clustered using kmeans method into families and the parameter K used was around 300 on average for each part, depending on the number of data points in each part. In the end, a total of around 60000 chemical families were identified from the Pubchem compound library. And the performance of this newly clustered chemical families is still under evaluation. Preliminary results indicate that SVM using putative negatives generated by these new chemical families could scan the 29.7 million Pubchem compound library with a lower hit rate. But the performance of SVM in terms of specificity and sensitivity still needs to be fine-tuned.

In addition, SVM can be modified to accommodate the need for lower false hit rate through iteratively throwing away the non-support vector negatives and adding in new putative negatives. Through this method, SVM models are expected to better differentiate those compounds situated at the border of SVM hyperplane.

3.2 Hints of drug prolific regions and properties by clustering drugs in the target-specific chemical space

The above VS tools developed have a reasonably good yield to identify virtual hits from the large chemical libraries. But as reviewed in the introduction, from hit to lead and from lead to drugs, the drug discovery process is still lengthy and costly, especially with the high attrition rates in clinical trials. More methods were needed to shorten the process and to increase the success rate. A hierarchical clustering method was proposed to cluster known drugs in the target-specific chemical space. This chemical space is spanned by the compounds from large chemical libraries whose structures are similar to drugs and inhibitors directed at a specific target. The clustering of known drugs will aid in the search of potential targeted drugs with good structure scaffold and optimal drug properties that have higher chance to enter clinical trials and ultimately into the market. Due to time constraint, this is only a preliminary investigation of possible drug prolific regions indicating privileged drug-like structure scaffolds and possible drug-like property rules that differentiate drugs from the inhibitors with similar structure scaffolds. The main focus of this section will be to present the workflow of applying the hierarchical clustering method to cluster drugs and inhibitor in the target-specific chemical space of structurally similar bioactive and non-active compounds.

3.2.1 Data collection and method

Ten therapeutic targets involved in various diseases were chosen for comprehensive coverage of different target types, including kinases (ABL1, B-Raf, FLT3, mTOR, SRC), G-protein coupled receptors (Beta-2 adrenoreceptor(B2AR), Dopamine D1 receptor (DA1R)), anti-HIV target (HIV reverse transcriptase) and other classical therapeutic targets (ACE and COX2). The 2D structures and relevant information of their inhibitors were obtained from chembl database, with IC50/Ki/EC50 value less than 10uM. And the structurally similar bioactive compounds were also obtained from Chembl database, defined as Tanimoto coefficient > 0.9 against any of the inhibitors or drugs to that target. In the same way, the structurally similar non-bioactive compounds were obtained from Pubchem database, with Tanimoto coefficient score > 0.9 against any of the inhibitors or drugs to that target and structurally similar approved drugs directed at other targets were collected from TTD. The statistics of drugs, inhibitors, similar other approved drugs, similar bioactive Chembl compounds and similar non-bioactive Pubchem compounds are listed in **Table 3.3**.

Table 3. 3 Overall statistics of drugs, inhibitors, structurally similar approved drugs directed to other drugs, similar bioactive Chembl compounds and similar non-bioactive Pubchem compounds to be clustered.

Targets	Drugs	Inhibitors	Other approved drugs	Similar Cpds from chembl	Similar Cpds from pubchem
ABL1	13	791	20	5529	28130
ACE	15	659	22	5259	104838
B2AR	20	1162	78	8048	117943
B-Raf	9	413	6	1138	4128
COX2	37	1917	54	11820	89985
DA1R	24	594	136	7235	45379
FLT3	16	939	19	7401	51665
HIVRT	12	1810	41	9954	55903

mTOR	14	931	15	1796	19854
SRC	7	2038	48	12365	87246

Each compound was represented by their Pubchem fingerprint, which is an 881 bit binary substructure fingerprint calculated from 2D structures. Each bit shows a Boolean determination of the presence of, for example, an element count, a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure.

And the similarity is defined by Tanimoto coefficient calculated from the following formula.

$$\text{Tanimoto} = AB / (A + B - AB)$$

Where:

Tanimoto is the Tanimoto score, a fraction between 0 and 1.

AB is the count of bits set after bit-wise & of fingerprints A and B

A is the count of bits set in fingerprint A

B is the count of bits set in fingerprint B

Hierarchical clustering is a clustering method that groups data over different levels by creating a cluster tree. It could be implemented by merging smaller clusters from bottom up into larger ones, or by splitting large clusters into smaller ones from top down. The computational complexity of the top down approach with an exhaustive search is $O(2^n)$, which makes it impossible to be applied to large data sets. The bottom up agglomerative approach with a complexity of $O(n^3)$ was used in this work. The cluster tree is a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. The level of clustering can be chosen according to the problem under study. Hence, this clustering method was used in our study so that we could determine the level of clustering to best fit the idea of compound scaffold.

In order to decide which clusters should be combined, a measure of dissimilarity

between sets of compounds is required. Accordingly, a distance matrix, which measures the distance between pairs of compounds, and a linkage criterion which specifies the dissimilarity of sets as function of the pairwise distance of compounds in the set were used. The distance between pairs of compounds was defined to be 1-tanimoto coefficient. Different linkage criterion, namely complete, single and average linkages were all used for comparison and evaluation.

The hierarchical clustering was done using the matlab `linkage()` function. And the Pubchem fingerprints were calculated using the software PaDEL-descriptor (139). The results of hierarchical clustering were stored in a newick format using matlab `phytreewrite()` function, which were then used as inputs to draw graphic trees through the online service interactive tree of life (iTol) (140).

For each of the ten drug targets, hierarchical clustering method was applied and the resulting clustering tree was cut at Tanimoto distance of 0.6. This split of overall tree into sub-trees were necessary for display of the circular trees in iTol, due to its limited display capacity (maximum of 10000 compounds can be displayed in an iTol tree graph). In each sub-tree, drugs, inhibitors, other drugs, similar chembl compounds and similar pubchem compounds are colored differently.

Additional physiochemical properties were analyzed and labeled on the distribution graphs, such as potency, ligand efficiency, molecular weight and logP value of the compound whenever available. The half maximal inhibitory concentration (IC_{50}) was used to indicate the potency of a compound. Ligand efficiency (LE) is the measurement of binding energy of per non-hydrogen atom of the compound, which is related to IC_{50} . It is defined as the ratio of Gibbs free energy to the number of non-hydrogen atoms of the compound and can be transformed to the following formula(141).

$$LE = 1.4(pIC_{50})/N,$$

where N is the number of non-hydrogen atoms

and $pIC_{50} = -\log(IC_{50})$.

Molecular weight (MW) and logP were calculated from our own software (117). For easy view, the values of these properties were transformed and rounded. In particular, the potency value was transformed to pIC_{50} and rounded to the smallest integer not less than itself. LE was rounded and scaled to an integer accordingly. MW was classified into several ranges, <50, 50~150, 150~250, 250~350, and so on. LogP was rounded to the smallest integer not less than itself. The values of these properties were reflected in the heights of the labels to the outer layer of the circular tree using iTol and colored differently according to the type of the compound.

3.2.2 Preliminary results

Table 3.4 shows the statistics of compounds in each subtree by cutting the hierarchical clustering result at tanimoto distance 0.6 using the complete-linkage method. The subtrees containing drugs are highlighted in green and only a minority of subtrees for almost all the targets contain drugs. From the statistics, drugs have shown some clustering tendency. The possible drug concentrated regions could be further examined in the clustering tree results.

Four figures (**Figure 3.4, 3.5, 3.6, 3.7**) displaying the labeled clustering results of FLT3 subtree ID 10 are given examples for illustration. Their potency, ligand efficiency, calculated cLogP and molecular weight are displayed individually in each figure.

It looks like approved drugs may be possibly distinguished from other structurally similar inhibitors by either (1) higher potency than other inhibitors, (2) better LE,

LogP and MW with respect to other inhibitors, (3) location in the region where there are other approved drugs, (4) located in the region where there are no other bioactive compounds around, or where the drugs have better LE than other bioactive compounds (which may indicate that the drugs cannot bind to other targets efficiently to produce large enough negative effects and thus have a good safety profile).

All these are only preliminary conjectures made by looking at the target-specific drug distribution graphs. More rigorous examinations of such graphs are required to make justifiable statements based on more drug distribution graphs of more targets.

Table 3. 4 Statistics of drugs, inhibitors, structurally similar approved drugs directed to other drugs, similar bioactive ChEMBL compounds and similar non-bioactive Pubchem compounds in each subtree.

Target	Subtree ID	Drugs	Inhibitors	Other approved drugs	Similar Cpd from ChEMBL	Similar Cpd from Pubchem	Target	Subtree ID	Drugs	Inhibitors	Other approved drugs	Similar Cpd from ChEMBL	Similar Cpd from Pubchem
ABL1	2	8	171	498	1	3048	DA1R	8	6	129	1266	23	5232
ABL1	5	2	268	1320	7	3840	DA1R	15	4	20	214	25	473
ABL1	4	1	1	19	0	24	DA1R	13	3	17	376	1	4242
ABL1	9	1	72	418	4	920	DA1R	19	3	3	60	1	196
ABL1	10	1	195	263	1	380	DA1R	10	2	24	536	13	895
ABL1	1	0	3	6	0	0	DA1R	18	2	22	351	21	2982
ABL1	3	0	5	12	0	274	DA1R	6	2	3	68	3	408
ABL1	6	0	36	79	0	1738	DA1R	9	2	61	1643	4	14797
ABL1	7	0	7	1816	1	6930	DA1R	1	0	102	29	0	6
ABL1	8	0	1	95	5	6	DA1R	2	0	105	2047	23	12178
ABL1	11	0	9	82	1	739	DA1R	3	0	1	3	0	1
ABL1	12	0	12	873	0	9430	DA1R	4	0	6	4	0	1
ABL1	13	0	11	48	0	801	DA1R	5	0	5	52	7	380
B-Raf	3	5	129	316	0	1106	DA1R	7	0	1	4	0	62
B-Raf	7	3	230	373	6	1909	DA1R	11	0	1	2	1	120

B-Raf	6	1	0	78	0	105	DA1R	12	0	1	2	1	53
B-Raf	1	0	4	41	0	159	DA1R	14	0	14	161	2	483
B-Raf	2	0	25	194	0	364	DA1R	16	0	25	222	8	2040
B-Raf	4	0	1	85	0	270	DA1R	17	0	51	45	3	131
B-Raf	5	0	14	1	0	0	DA1R	20	0	3	150	0	699
B-Raf	8	0	10	50	0	215	COX2	2	5	17	212	2	5435
FLT3	10	8	216	1676	5	5873	COX2	14	5	36	371	5	3236
FLT3	16	3	205	1171	4	5868	COX2	25	4	208	334	1	1130
FLT3	17	2	102	71	0	271	COX2	8	3	38	183	1	2163
FLT3	1	1	47	310	0	458	COX2	15	3	38	123	1	1676
FLT3	4	1	42	766	3	3980	COX2	16	3	103	5347	6	28828
FLT3	9	1	45	570	3	14950	COX2	24	3	154	803	3	11048
FLT3	2	0	2	19	0	0	COX2	21	2	94	158	1	1031
FLT3	3	0	1	0	0	0	COX2	26	2	19	151	0	2949
FLT3	5	0	19	57	1	108	COX2	3	1	99	135	0	284
FLT3	6	0	20	64	0	194	COX2	6	1	22	119	4	2846
FLT3	7	0	43	115	0	1710	COX2	12	1	181	181	0	610
FLT3	8	0	18	222	0	5799	COX2	17	1	86	199	1	936
FLT3	11	0	2	3	0	10	COX2	22	1	44	134	2	1321
FLT3	12	0	1	2	0	5	COX2	23	1	156	242	3	1207
FLT3	13	0	66	635	1	1435	COX2	28	1	28	480	2	4664
FLT3	14	0	69	165	1	671	COX2	1	0	167	369	2	1712
FLT3	15	0	20	156	0	3107	COX2	4	0	17	86	0	877
FLT3	18	0	15	1379	1	7219	COX2	5	0	37	126	0	1067

FLT3	19	0	6	20	0	7	COX2	7	0	6	537	0	812
mTOR	18	5	41	190	3	1125	COX2	9	0	1	9	0	203
mTOR	13	3	263	335	0	1504	COX2	10	0	26	103	0	201
mTOR	1	2	75	112	0	621	COX2	11	0	63	212	1	4967
mTOR	11	2	10	31	0	306	COX2	13	0	28	41	0	98
mTOR	6	1	66	62	0	1715	COX2	18	0	111	88	0	1219
mTOR	10	1	0	285	4	467	COX2	19	0	5	1	0	2
mTOR	2	0	1	0	0	2	COX2	20	0	1	1	0	3
mTOR	3	0	25	1	0	52	COX2	27	0	8	120	0	677
mTOR	4	0	2	31	0	1740	COX2	29	0	59	72	12	537
mTOR	5	0	323	466	6	2809	COX2	30	0	19	839	7	7867
mTOR	7	0	0	0	0	2147	COX2	31	0	46	44	0	379
mTOR	8	0	1	0	0	4198	HIVRT	29	3	388	867	6	8266
mTOR	9	0	2	7	0	130	HIVRT	27	2	117	418	1	254
mTOR	12	0	1	16	0	10	HIVRT	34	2	42	293	13	727
mTOR	14	0	7	5	0	4	HIVRT	3	1	163	258	0	1297
mTOR	15	0	108	85	1	276	HIVRT	11	1	76	146	0	382
mTOR	16	0	5	170	1	221	HIVRT	16	1	25	225	0	435
mTOR	17	0	1	0	0	2527	HIVRT	18	1	16	257	4	754
SRC	2	2	318	824	0	1993	HIVRT	24	1	2	50	1	134
SRC	9	1	0	35	0	102	HIVRT	1	0	62	195	0	233
SRC	15	1	0	79	0	105	HIVRT	2	0	5	11	0	24
SRC	17	1	555	2295	8	6134	HIVRT	4	0	27	47	0	1333
SRC	18	1	439	1383	6	7927	HIVRT	5	0	15	94	3	125

SRC	19	1	107	314	1	778	HIVRT	6	0	9	46	0	107
SRC	1	0	7	29	0	632	HIVRT	7	0	61	199	0	751
SRC	3	0	10	51	1	230	HIVRT	8	0	22	92	1	611
SRC	4	0	151	668	4	703	HIVRT	9	0	169	415	0	3292
SRC	5	0	14	24	0	49	HIVRT	10	0	18	35	0	339
SRC	6	0	10	217	0	12969	HIVRT	12	0	44	60	0	131
SRC	7	0	1	1	0	114	HIVRT	13	0	10	9	0	13
SRC	8	0	126	704	4	1563	HIVRT	14	0	8	12	0	47
SRC	10	0	21	47	0	1531	HIVRT	15	0	3	5	0	77
SRC	11	0	2	37	0	591	HIVRT	17	0	19	161	1	573
SRC	12	0	45	49	2	299	HIVRT	19	0	71	264	1	1466
SRC	13	0	20	2490	8	9997	HIVRT	20	0	2	2	0	10
SRC	14	0	3	247	5	466	HIVRT	21	0	1	1	1	9
SRC	16	0	11	22	0	595	HIVRT	22	0	1	0	1	17
SRC	20	0	31	127	0	1220	HIVRT	23	0	30	207	0	816
SRC	21	0	125	2632	2	38225	HIVRT	25	0	105	274	0	98
SRC	22	0	38	46	0	98	HIVRT	26	0	32	23	0	193
SRC	23	0	4	44	7	925	HIVRT	28	0	15	50	0	114
B2AR	1	7	62	576	18	5137	HIVRT	30	0	16	368	1	8333
B2AR	3	4	276	272	0	1892	HIVRT	31	0	28	47	0	1532
B2AR	10	2	29	122	4	1585	HIVRT	32	0	46	191	0	1158
B2AR	12	2	316	845	7	2019	HIVRT	33	0	43	830	5	2802
B2AR	15	2	44	706	15	10298	HIVRT	35	0	48	158	0	219
B2AR	17	2	113	925	1	16568	HIVRT	36	0	71	3644	2	19231

B2AR	16	1	47	280	0	891	ACE	15	11	302	1816	13	32387
B2AR	2	0	34	871	0	10182	ACE	6	1	73	1681	3	33245
B2AR	4	0	1	2	2	13	ACE	8	1	23	58	0	655
B2AR	5	0	16	393	0	2090	ACE	12	1	55	106	0	1256
B2AR	6	0	5	117	0	268	ACE	14	1	41	104	0	4765
B2AR	7	0	17	73	0	448	ACE	1	0	10	5	0	9
B2AR	8	0	4	23	0	296	ACE	2	0	50	124	0	2502
B2AR	9	0	7	124	8	196	ACE	3	0	4	12	1	60
B2AR	11	0	6	127	2	971	ACE	4	0	22	24	1	428
B2AR	13	0	4	28	0	183	ACE	5	0	3	46	3	817
B2AR	14	0	9	9	0	4	ACE	7	0	14	187	0	7175
B2AR	18	0	6	24	0	21	ACE	9	0	1	1	0	147
B2AR	19	0	26	1544	11	53160	ACE	10	0	7	7	0	70
B2AR	20	0	28	315	0	3247	ACE	11	0	7	9	0	35
B2AR	21	0	63	113	0	412	ACE	13	0	40	1078	1	21253
B2AR	22	0	44	340	8	4607	ACE	16	0	7	1	0	34
B2AR	23	0	5	219	2	3455							

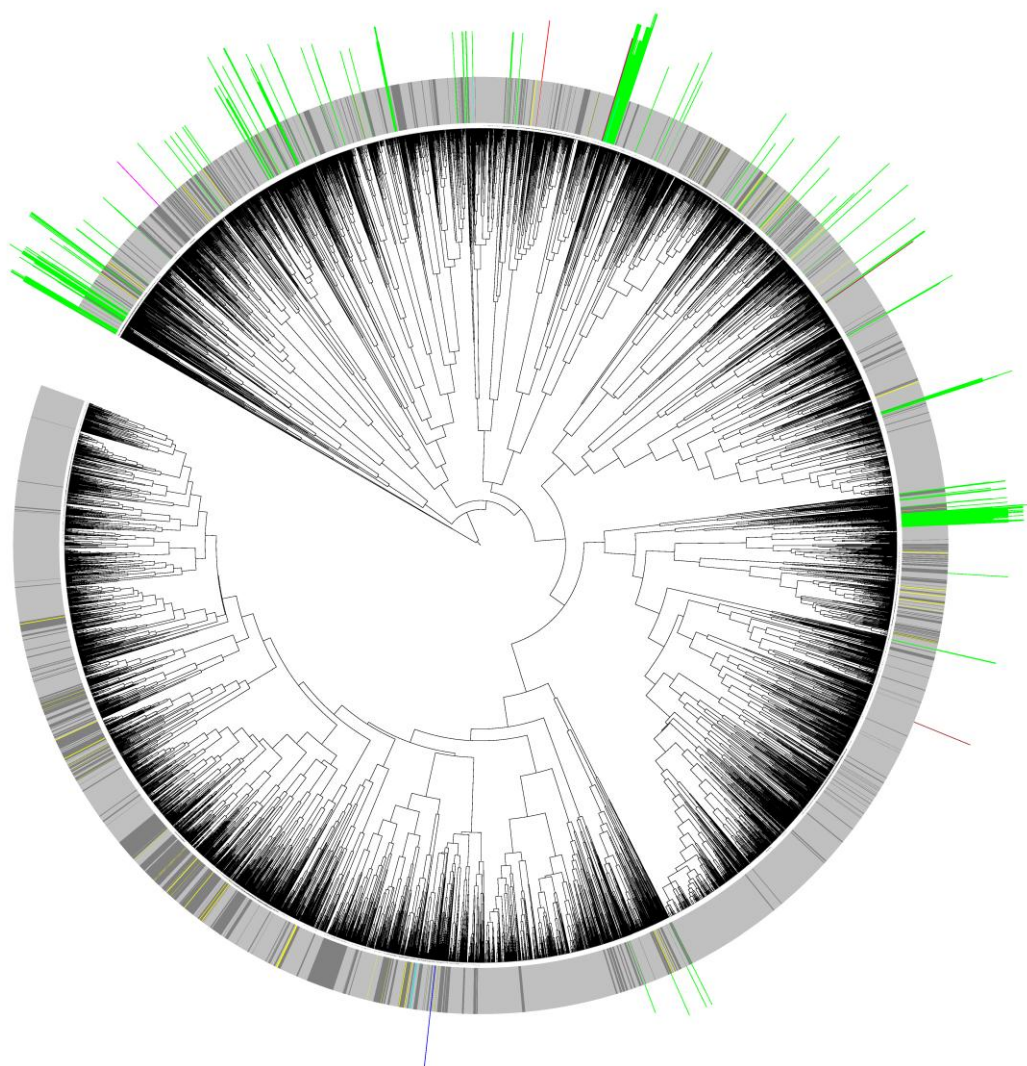


Figure 3. 4 Distribution graph of FLT3 subtree ID 10, labelled according to potency values. The labels are colored as follows: red for Approved drug, purple for Phase III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for inhibitors, grey for similar ChEMBL compounds, pale grey for similar Pubchem compounds.

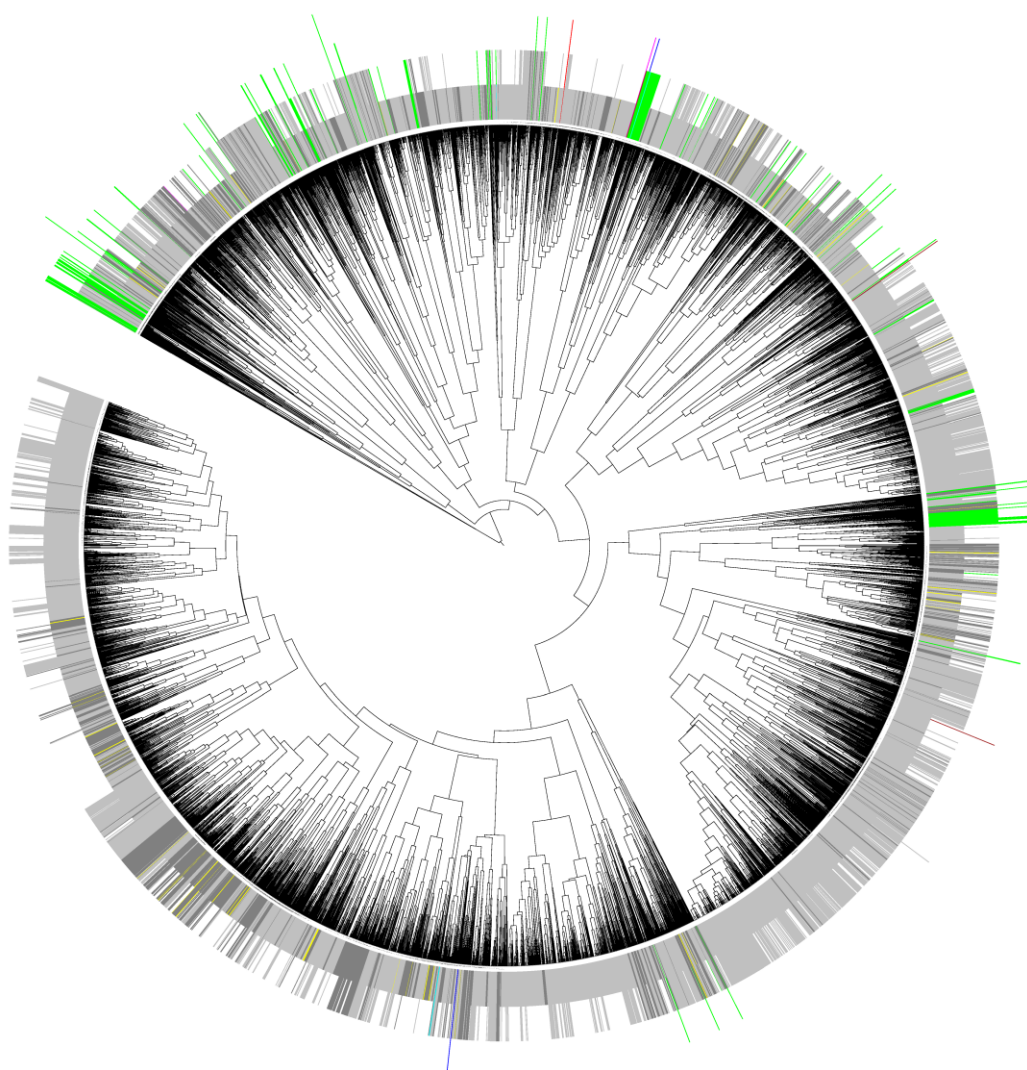


Figure 3. 5 Distribution graph of FLT3 subtree ID 10, labelled according to ligand efficiency values. The labels are colored as follows: red for Approved drug, purple for Phase III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for inhibitors, grey for similar ChEMBL compounds, pale grey for similar PubChem compounds

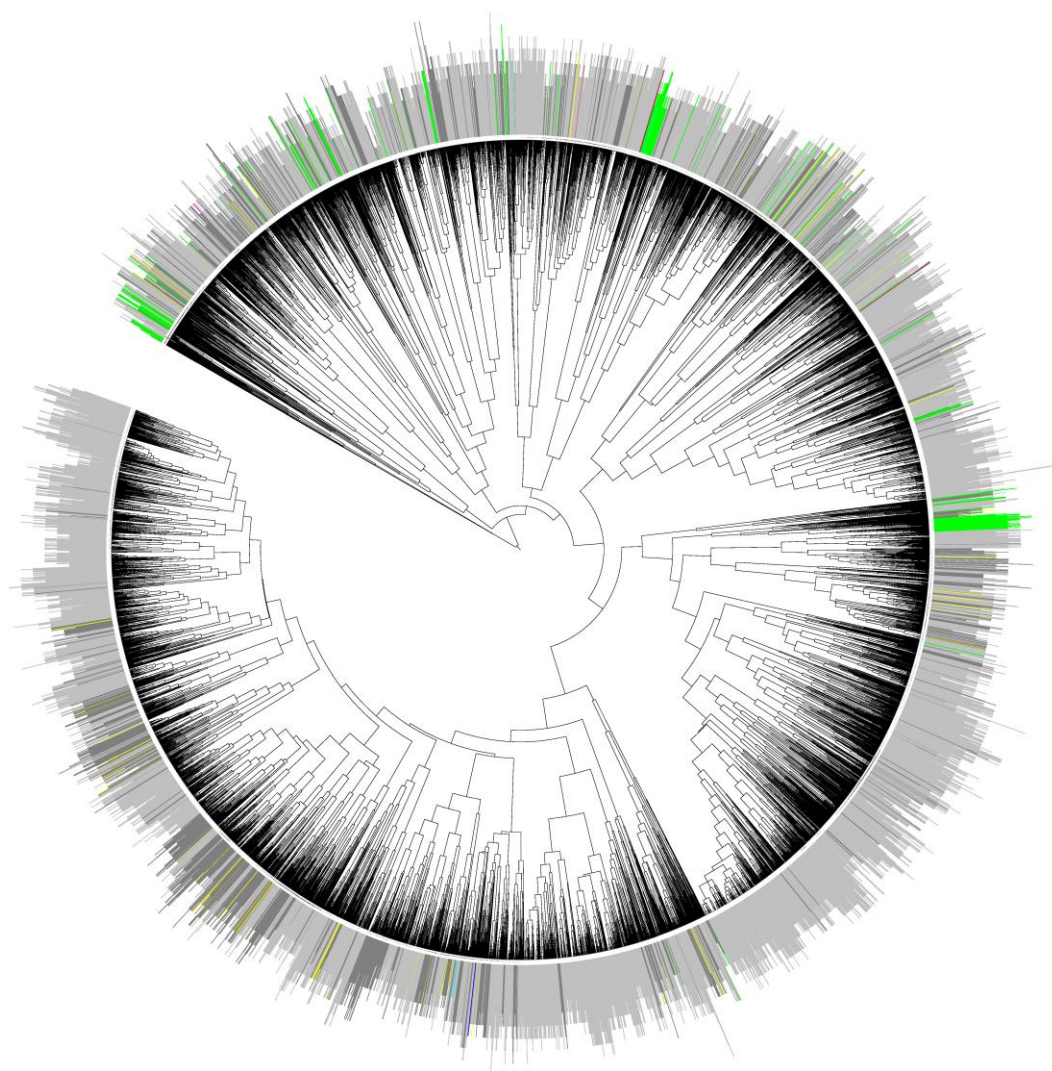


Figure 3. 6 Distribution graph of FLT3 subtree ID 10, labelled according to the calculated clogP values. The labels are colored as follows: red for Approved drug, purple for Phase III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for inhibitors, grey for similar ChEMBL compounds, pale grey for similar Pubchem compounds

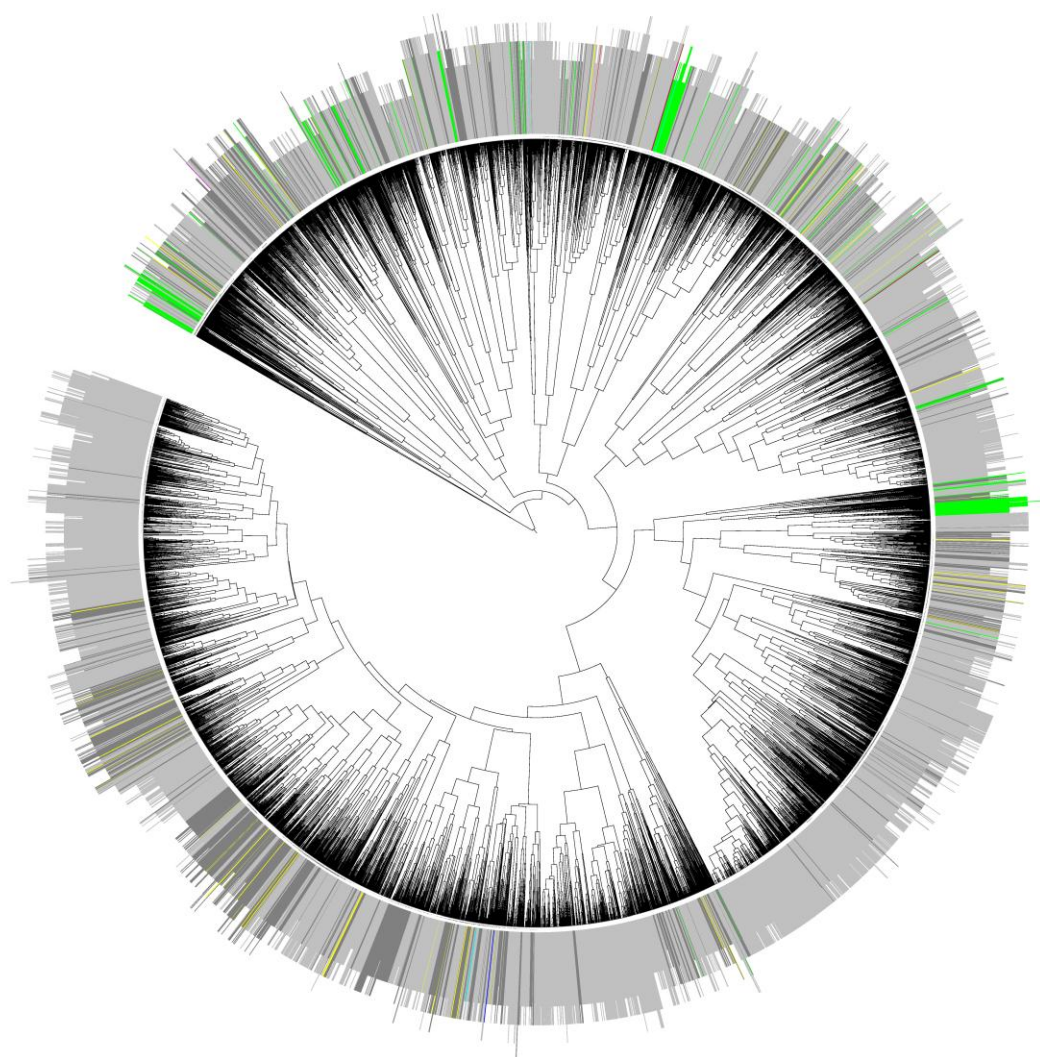


Figure 3. 7 Distribution graph of FLT3 subtree ID 10, labelled according to molecular weight. The labels are colored as follows: red for Approved drug, purple for Phase III drug, pink for Phase II drug, blue for Phase I drug, cyan for other drugs, green for inhibitors, grey for similar ChEMBL compounds, pale grey for similar Pubchem compounds

Chapter 4. Specific multi-target modes identified by analysing synergistic natural product combination

In the previous chapter, various machine learning methods and clustering method to learn from the structures and properties of known drugs and inhibitors for virtual screening and drug design have been presented. While most of the drugs and inhibitors are of synthetic origins based on the success of combinatorial chemistry, a significant portion of drugs and inhibitors nowadays are still derived from nature.

Combinatorial chemistry has been an important source of creating new chemical entries over the past decades. However, despite the explosion of synthetic chemical space, people are increasingly concerned about the low yield of chemical synthesis to generate lead compounds. The low drug productivity has directed people's interest to natural products as drug discovery sources. Natural products derived from plants, microbial and marine species have a rich diversity of structures and good therapeutic properties. They have performed well as a major source of therapeutics for infectious diseases, lipid disorders, immunomodulation and cancer. Out of the 175 FDA approved small molecule anticancer drugs, 85 of them are either natural products or natural derivatives(142).

The information contained in natural products is of great use in drug discovery. In particular, natural product (NP) combinations, in many cases as combinations of

whole herbs or herbal extracts, may be useful sources for developing new drug combinations based on their novel multi-targeted mechanisms and potentially give clues to the design of multi-targeted drug combinations.

In this chapter, synergistic natural product combinations will be analysed systematically. Four important questions need to be answered for assessing the possible contribution of synergism to the therapeutic efficacies: what are the gaps between the potencies of the typically studied bioactive NPs and those of drugs, whether synergistic combination of sub-potent NPs can sufficiently enhance their collective potencies to reach drug potency levels, and at what odds and by what molecular modes such NP combinations can be assembled.

The first question was studied by analyzing the literature-reported cell-based potencies of 190 approved drugs and 1378 NPs of anticancer and antimicrobial classes. Potencies derived from cell-based assays were used instead of target-based and in-vivo assays for several reasons. To a certain extent, cell-based assays can predict some level of in-vivo activities (143, 144) and these assays have been successfully used for discovering therapeutic agents that entered advanced development stages (145). Within the same disease classes, cell-line assays are more mutually comparable and better reflecting overall effects of targeted actions and intracellular bioavailability than target-based assays. The number of NPs with cell-based potency data is significantly higher than those with in-vivo data. The

anticancer and antimicrobial classes were focused because of the availability of statistically significant number of cell-based activity data, the relatively comparable bioassays than some other therapeutic classes, and the relevance to our NP combination studies (67% of our studied synergistic NP combinations are from these two classes).

The second question was addressed by evaluating 124 literature-reported synergistic combinations of 158 NPs with cell-based activity data available for all of the constituents both in individual and in the respective combination. These data are necessary for deriving combination index (CI) and dose reduction index (DRI, ratio of the effective dose in individual and in combination) to allow rigorous and quantitative evaluation of synergistic effects (146). The third question was probed by analyzing 122 molecular interaction profiles (MIPs) in 19 NP combinations with potencies enhanced to drug levels or by over 10-fold. These MIPs are linked to the potency-enhancing synergistic actions of these NP combinations, and their analysis reveals general molecular modes for significantly enhancing potency via collective modulation of specific targets and their regulators and effectors, and the pharmacokinetics of the active NP ingredients(73, 77).

While these 122 MIPs have been individually reported in the literatures, to the best of our knowledge, few of them have been collectively analyzed for probing potency enhancing molecular modes in specific NP combinations. It is cautioned that,

although connections can be made between these MIPs and the synergistic potency-enhancing modes of the NP combinations, many of these interconnections are much more complicated than those analyzed here, and their activities are highly dynamic (147-149). The activation and the activity levels of these interconnections may be influenced by genetic variations (150), environmental factors (151), host's behavior (152), and therapeutic scheduling (153). Therefore, the use of these interconnections should be more appropriately viewed as a start to a more comprehensive analysis of the potency-enhancing modes in NP combinations.

4.1 Method

Experimentally determined cell-based inhibitory activities of anticancer and antibacterial drugs and NPs were searched from the Pubmed database (94) by using keyword combinations of 'drug', 'natural product', 'herb', 'medicinal plant', 'extract', 'ingredient', 'GI50', 'IC50', 'MIC', "activity", 'cell-line', and 'in vitro'. Cell-based inhibitory activities of 88 anticancer drugs and 102 antimicrobial drugs were obtained from the literatures and the NCI standard agent database. The approval status of these drugs was further checked against the drug data in the Therapeutic target database. Cell-based inhibitory activities of 1378 anticancer and antimicrobial NPs and 99 antimicrobial NP extracts were obtained from the literatures. These activities are typically given as GI50 or IC50 values against cancer cell-lines or MIC values against microbial cells. For drugs and NPs with multiple

potency data, the best potency was selected.

Literature-reported synergistic NP combinations were searched from the Pubmed database (94) by using keyword combinations ‘natural product’, ‘herb’, ‘medicinal plant’, ‘extract’, ‘ingredient’, ‘synergistic’, ‘synergy’, ‘synergism’, ‘synergize’, and ‘potentiate’. The full reports of the searched articles were evaluated to select those synergistic NP combinations with the experimental cell-based activities available for all constituent NPs both as individual and in the respective combination. Although many NP combinations have been reported to show synergism (71, 73, 85, 86), only 124 synergistic combinations of 158 NPs are with available cell-based activity data for enabling the computation of CI and DRI values. The cell-based activities of the constituent NPs in some of these combinations are given in terms of the percent inhibitory rates at particular concentrations. The CIs and DRIs of these combinations were computed by using the median effect equation, the multiple drug effect equation, and the combination index theorem outlined by Chou (146).

$$\begin{aligned}(DRI)_1 &= \frac{(D_x)_1}{(D)_1} \\ (DRI)_2 &= \frac{(D_x)_2}{(D)_2} \\ CI &= \frac{1}{(DRI)_1} + \frac{1}{(DRI)_2}\end{aligned}$$

Where D = Dose, CI: combination index, DRI: dose-reduction index

$(D_x)_1$: dose of drug 1 alone to achieve 50% inhibitory effects.

$(D)_1$: dose of drug 1 used in the combination to achieve 50% inhibitory effects.

4.2 Results and discussion

4.2.1 Comparison of the potencies of natural products and drugs in cell-based assays

Drug potency is context dependent, varying with assay, target and technology. Previous analysis of existing drugs has suggested that drugs in cell-based assays typically exhibit potencies of $\leq 1\mu\text{M}$ (154). Hence, in our analysis, drug potency levels for anticancer and antimicrobial classes were tentatively taken as $\text{GI}_{50}/\text{IC}_{50} \leq 1\mu\text{M}$ and $\text{MIC} \leq 1\mu\text{g/mL}$, which are satisfied by 76% anticancer and 86% antimicrobial drugs respectively. It is noted that, in some cases, drug efficacy is not only determined by cell-based activities. Some drugs sub-potent in cell-based assays are nonetheless clinically efficacious by such additional mechanisms as immuno- and hormone modulations (155, 156). While drug potency levels can be more rigorously defined by consideration of these mechanisms, the possible contribution of these mechanisms has been studied for few drugs and NPs sub-potent in cell-based assays. Therefore, it is more practically feasible to tentatively focus on cell-based activities to enable potency analysis of statistically significant number of drugs and NPs.

Figure 4.1 and **4.2** show the potency distribution profiles of 88 and 650 anticancer drugs and NPs, and those of 102, 609 and 99 antimicrobial drugs, NPs and NP extracts respectively. The median potencies of anticancer ($\text{GI}_{50}/\text{IC}_{50}=28\text{nM}$) and antimicrobial ($\text{MIC}=0.12\mu\text{g/mL}$) drugs are 214-fold and 104-fold higher than those

of anticancer ($GI_{50}/IC_{50}=6\mu M$) and antimicrobial ($MIC=12.5\mu g/mL$) NPs. Overall, 25% of the anticancer and 10% of the antimicrobial NPs reach drug potency levels, and additional 33% of the anticancer and 37% of the antibacterial NPs are within 10-fold range of drug potency levels ($1\mu M < GI_{50}/IC_{50} \leq 10\mu M$, $1\mu g/mL < MIC \leq 10\mu g/mL$). There is a small pool of drug level potent NPs (10-25%). It is noted that a significantly larger pool of NPs (47-58%) may be explored for designing NP combination therapies if synergistic combinations of >10-fold potency enhancement can be assembled at reasonable probabilities. The potencies of the NP extracts are mostly 100-1,000 folds lower than those of individual NPs, partly because the active constituents in each NP extract typically constitute a small percent of the contents (157). Because of their 100-1,000 fold lower potencies, NP extracts have been typically prescribed in g/kg (158) (159) instead of the usual mg/kg for individual drugs.

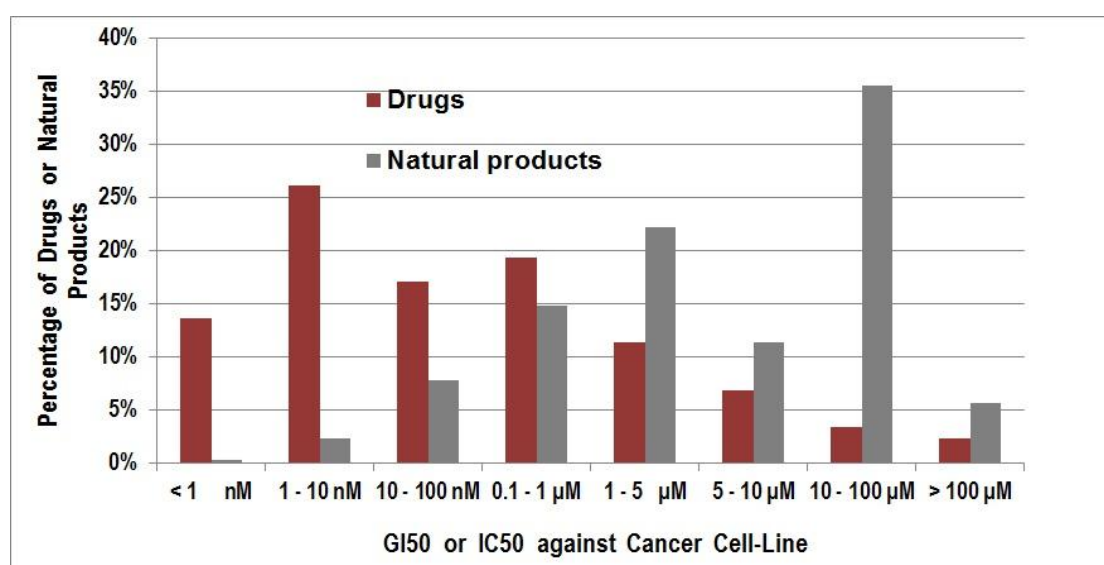


Figure 4. 1 Potency distribution profiles of 88 and 650 anticancer drugs and natural

products.

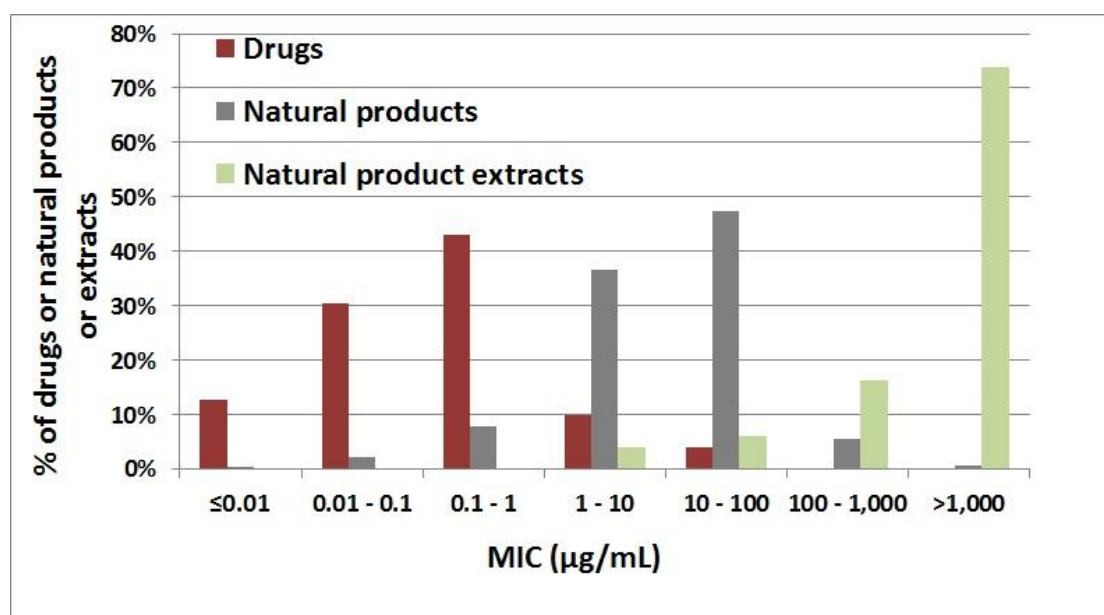


Figure 4. 2 Potency distribution profiles of 102, 609 and 99 antibacterial drugs, natural products (NPs) and NP extracts.

4.2.2 Synergistic natural product combinations

Based on Chou's method (146), the levels of synergism in these NP combinations (**Figure 4.3**) were categorized into very strong synergism ($CI < 0.1$), strong synergism ($CI = 0.1 - 0.3$), synergism ($CI = 0.3 - 0.7$), moderate synergism ($CI = 0.7 - 0.85$), slight synergism ($CI = 0.85 - 0.90$), nearly additive ($CI = 0.90 - 1.10$), slight antagonism ($CI = 1.10 - 1.20$), and moderate antagonism ($CI = 1.20 - 1.45$) respectively. Overall, 24% and 34% of the combinations are at the strong/very strong synergism and synergism levels, indicating that highly synergistic combinations can be formed at fair probabilities. **Figure 4.4** shows the potency improvement profile of the NPs in these combinations, in which 4% and 19% of the NPs exhibit >100-fold and

10-100 fold potency improvement respectively. This suggests that >10-fold potency improvement is achievable at moderate probabilities. These combinations are mostly composed of sub-potent NPs. There are only 6 potent NPs, and 1 and 3 combinations fully and partially composed of potent NPs. Synergism elevates the group potencies (potencies of all components) of 5 fully sub-potent and 2 partially sub-potent combinations to drug levels, and lifts the potency of 4 NPs in another 3 sub-potent NP combinations to drug levels. Overall, the potencies of 22 (14.4%) sub-potent NPs and group potencies of 7 (5.6%) sub-potent combinations are enhanced to drug levels, suggesting that the individual and group potencies of sub-potent NPs can be raised to drug levels at moderate and low probabilities respectively.

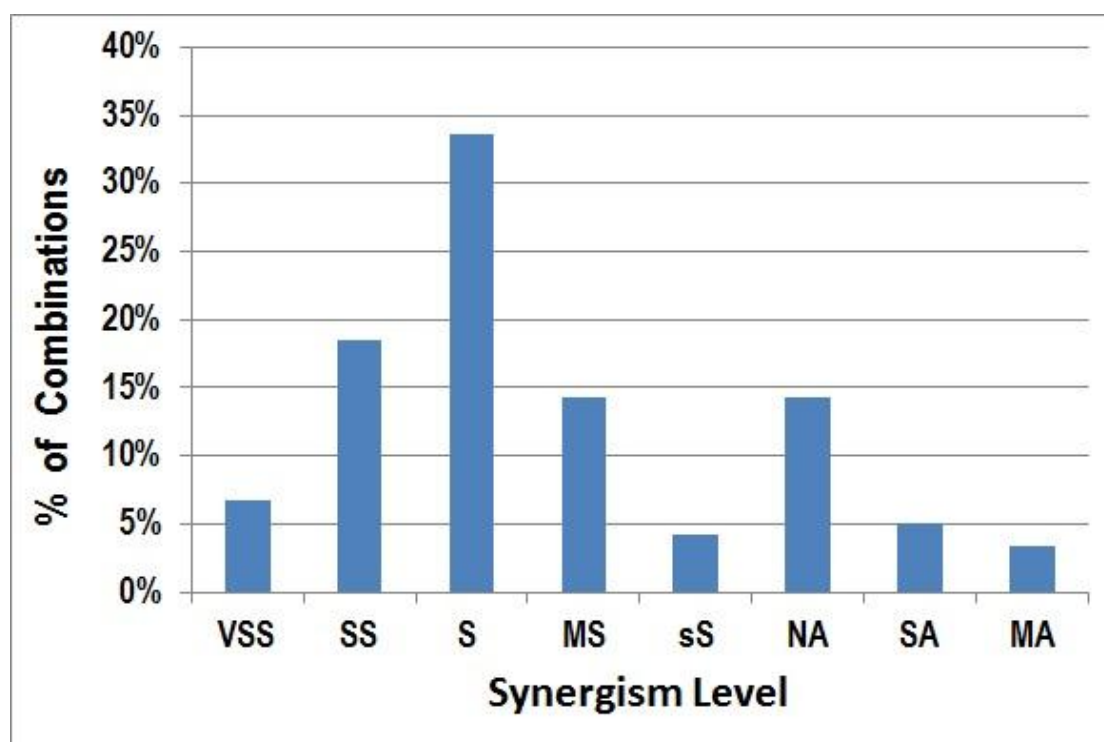


Figure 4. 3 Synergism level of 124 synergistic NP combinations. VSS, SS, S, MS, sS: very strong, strong, normal, moderate, slight synergism, NA: nearly additive, SA, MA: slight, moderate antagonism.

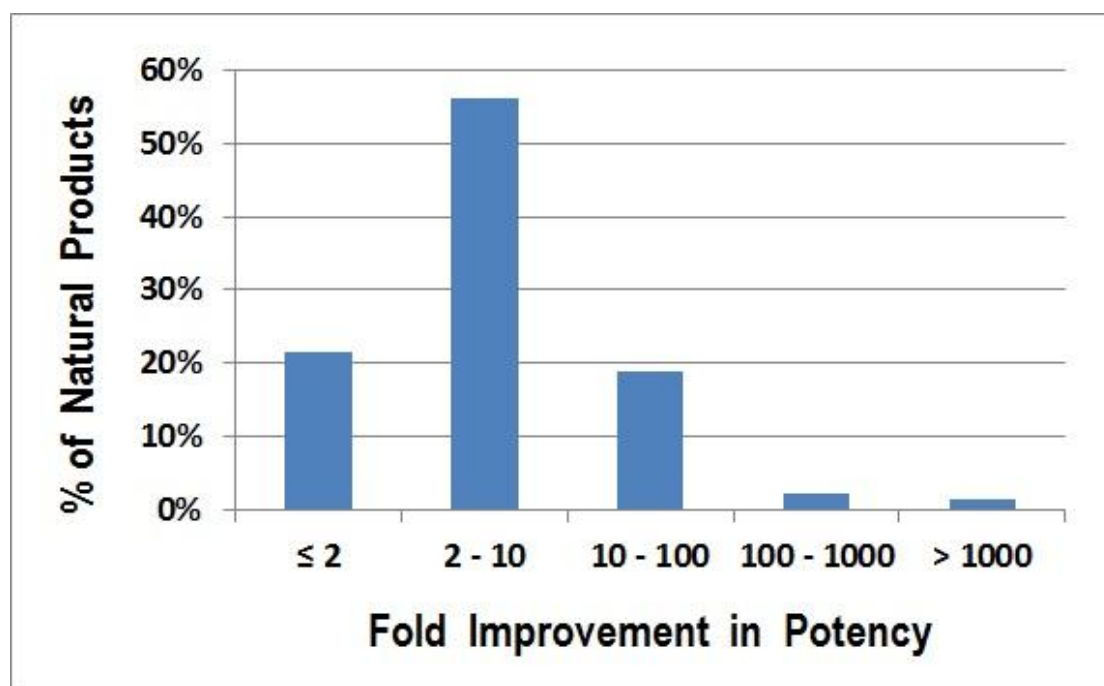


Figure 4. 4 The potency improvement profile of the constituent NPs.

4.2.3 Potency enhancing molecular modes of natural product combinations

The molecular mechanisms of synergism of drug combinations (77) and NP combinations (73) can be studied from their MIPs. We conducted comprehensive literature search for identifying the targets and synergism-related MIPs of three NP combinations with collective potencies improved to drug levels, which identified 11 targets related to the reported therapeutic effects of these combinations and 72 MIPs likely contributing to the potency-enhancing modes (**Supplementary Table S4.1**). The targets and potency-enhancing MIPs of two of the NP combinations are also summarized in **Table 4.1** and **4.2**. Specific potency-enhancing molecular modes were identified. The potencies of the principal NP in these combinations are at or

near drug potency levels ($IC_{50}=0.8-1.1\mu M$, $0.94\mu g/mL$) probably due in part to the multi-target activities of each principal NPs (2, 4, 5 targets respectively). Network models and activity assays have shown that weak inhibition of multiple targets in related pathways may be more efficient than strong inhibition against a single target (160, 161). The potencies of the companion NPs are substantially weaker ($IC_{50}=1.7-656\mu M$, $5.07-251\mu g/mL$). The potencies of all NPs in these combinations are significantly enhanced (mostly by >10 -fold) by multi-target actions in modulating multiple regulators, partners and effectors of the primary targets of the principal and the active companion NPs (complementary actions), elevating intra-cellular bioavailability of the principal and the active companion NPs, and antagonizing the processes counteractive to the therapeutic effects of the principal and the active companion NPs (anti-counteractive actions).

Table 4. 1 The targets and potency-enhancing synergistic molecular modes of the anticancer combination of Tetraarsenic tetrasulfide, Indirubin, and Tanshinone IIA (anticancer synergism reported in literature(162))).

Natural Product [Role in Combination] (Individual Potency) { Dose Reduction Index}	Target, Therapeutic Effect or Response (reference in Pubmed ID)	Effect type	Potency-Enhancing Synergistic Modes (reference in Pubmed ID)	Type of Synergism
Tetraarsenic tetrasulfide [Principal] (1.1uM) {6.88}	Degraded PML-RAR to produce anticancer effect (18344322)	Growth inhibition,	Indirubin blocked RAR-STAT3 crosstalk (14959844) by reducing JAK/STAT3 signaling (21207415). Tanshinone IIA reduced RAR (12069693) by hindering AR (22175694, 22281759, 21997969). These complement tetraarsenic tetrasulfide's action on RAR	Complementary action
	Down-regulated CDK2 in NB4 and NB4-R2 cells (18344322)	Cell cycle regulation	Indirubin inhibited and reduced CDK2 (18344322) to complement tetraarsenic tetrasulfide's action on CDK2	Complementary action
	Upregulated RING-type E3 ligase c-CBL and degraded BCR-ABL (21118980)	Growth inhibition		

	Transported into tumor cells by AQP9 (18344322)	Intracellular bioavailability	Indirubin and Tanshinone IIA upregulated AQP9 (18344322) to promote Tetraarsenic tetrasulfide's cell entry	Intracellular bioavailability enhancement
	RAR α reduction downregulated P53 and elevated Bcl-2 (10675490) to reduce apoptosis	Counteractive action	Tanshinone IIA activated p53 signaling (21997969) to reduce this counteractive action	Anti-counteractive action
Indirubin [Cooperative] (>3uM) {>9.38}	Inhibited and reduced CDK2 to produce anticancer effect (18344322)	Cell cycle regulation	Tetraarsenic tetrasulfide reduced CDK2 (18344322) to complement indirubin's action on CDK2	Complementary action
	Inhibited GSK3 to produce anticancer effect (21697283)	Growth inhibition		
	blocked VEGFR2 signaling (21207415) to reduce angiogenesis and apoptosis (14959844)	Growth, angiogenesis inhibition		
	Activated AhR (20951181) which activates RAR α (16480812) to promote cancer	Counteractive action	Tetraarsenic tetrasulfide degraded PML-RAR (18344322) to alleviate this counteractive action	Anti-counteractive action
Tanshinone IIA [Cooperative] (>3uM) {>9.38}	Increased Bax/Bcl-2 ratio, caspase 3, reduced Bcl-2, mitochondrial membrane potential, MMPs, to promote apoptosis (21472292, 22002472, 22126901)	Apoptosis		

	Activated p53 signaling to promote anticancer effect (21997969)	Cell cycle regulation, apoptosis		
	Upregulated pP38 to enhance apoptosis (21165580)	Apoptosis		
	Reduced HER2, NF- κ Bp65, RAR α activities (17451432) to promote anticancer effect (22246196),	Apoptosis, growth inhibition,		
	Reduced and antagonized AR and induced apoptosis (22175694, 22281759, 21997969)	Growth inhibition		
	pP38 upregulation (21165580) activated RAR α (19078967, 20080953) to promote cancer	Counteractive action	Tetraarsenic tetrasulfide degraded PML-RAR (18344322) to alleviate this counteractive action	Anti-counteractive action
	Upregulated efflux transporters to promote Tanshinone IIA (a Pgp substrate) efflux (17504222, 20821829)	Intracellular bioavailability	Indirubin inhibit certain efflux pumps (20380543) which may reduce the efflux of Tanshinone IIA	Intracellular bioavailability enhancement

Table 4. 2 The targets and potency-enhancing synergistic molecular modes of the anti-rotavirus combination of Theaflavin, Theaflavin-3-monogallate, Theaflavin-3'-monogallate, and Theaflavin-3,3' digallate (anti-rotavirus synergism reported in literature (163)).

Natural Product [Role in Combination] (Individual Potency) { Dose Reduction Index}	Target, Therapeutic Effect or Response (reference in Pubmed ID)	Effect type	Potency-Enhancing Synergistic Modes (reference in Pubmed ID)	Type of Synergism
Theaflavin [Principal] (0.943ug/mL) {9.33}	Reduced JNK and P38 phosphorelation (21184129, 22111069) to block JNK and p38 mediated viral replication	Viral replication inhibition	Other 3 components block the redundant Cox2 and ERK viral replication pathways to complement Theaflavin's activity	Complementary action
Theaflavin-3-monogallate [Cooperative] (251.39ug/mL) {2489}	Theaflavin-3-monogallate and theaflavin-3'-monogallate mixture downregulated Cox2 (11103814) to block Cox2 mediated viral replication and infection (15331705, 17555580)	Viral replication inhibition	All 4 components collectively cover 4 redundant viral replication pathways to complement Theaflavin-3-monogallate's activity	Complementary action
Theaflavin-3'-monogallate [Cooperative] (5.07ug/mL) {50.2}	Theaflavin-3-monogallate and theaflavin-3'-monogallate mixture downregulated Cox2 (11103814) to block Cox2 mediated viral replication and infection (15331705, 17555580),	Viral replication inhibition	All 4 components collectively cover 4 redundant viral replication pathways to complement Theaflavin-3'-monogallate's activity	Complementary action
Theaflavin-3,3' digallate [Cooperative] (5.51ug/mL) {54.6}	Reduced ERK phosphorelation (11511526) to block ERK mediated viral replication (17689685),	Viral replication inhibition	Other 3 components block the redundant JNK, P38 and Cox2 viral replication pathways to complement Theaflavin-3,3'	Complementary action

			digallate's activity	
	Blocked NFkB activation (16880762) to hinder NFkB and AkT mediated viral survival and growth (20392855)	Viral survival, growth inhibition		

Regulation of multiple regulators of the primary targets of principal NPs is important for elevating the collective potencies to drug levels. In two combinations, 6 and 13 regulators of the primary targets of the principal NPs are modulated. In the third combination, each constituent NP targets one or two of the four redundant processes to collectively achieve therapeutic effects. These multi-target potency-enhancing modes are consistent with the reports that weak inhibition of multiple targets in related pathways may be more efficient than strong inhibition of a single target(160, 161). In these combinations, complementary actions are achieved by modulating the expression, upstream regulators, crosstalk/redundant signaling, and substrates/effectors of the targets of individual NPs. Intra-cellular bioavailability of NPs are enhanced by inhibiting/downregulating efflux pumps and upregulating/activating cell-entry transporters. Anti-counteractive actions involve regulation of the pathways activated by the NPs that subsequently reduce the therapeutic effects of the NPs. Drug efficacies are reportedly reduced by network robustness (164), redundancy (18), crosstalk (19), and compensatory and neutralizing actions (20). Our revealed potency-enhancing molecular modes of synergistic natural products combinations are consistent with these literature-reported findings and provide clues for multi-target strategies in reducing these negative effects.

Additional potency-enhancing mechanisms were studied by analyzing 8 and 26 MIPs in 2 and 9 combinations with the potency of the principal NP enhanced

by >100-fold and 10-100 fold, and 16 MIPs of 5 combinations with the potency of a non-principal NP improved by >10-fold respectively. (**Supplementary Table S4.2, S4.3, S4.4**) The potency of individual NPs in 13 combinations is enhanced by a single mechanism: enhancement of the intra-cellular bioavailability of an active NP, which is an extensively-explored and effective potency-enhancing strategy for those NPs with hindered intra-cellular bioavailability. In addition to actions on efflux and cell-entry transporters, intra-cellular bioavailability of NPs can be enhanced by regulating their metabolism, disrupting membrane structures, and the use of pro-drug NPs of better cell-entry abilities, The potency of individual NPs in the remaining 3 combination is enhanced by complementary and anti-counteractive modes similar to those of the three NP combinations with potencies improved to drug levels.

Although the potencies of some of the individual NPs in these combinations are significantly improved, none is elevated to drug levels possibly due to low potencies of their principal NPs (44.6-800µg/mL with one exception) and modulation of few (60, 61) regulators of the targets of the principal NPs. The success rate of assembling sub-potent NPs into drug-level potent combinations may be significantly improved by careful selection of principal NPs of sufficient potency (e.g. potency <10µM) and the use of cooperative NPs that enhance the bioavailability and modulate the regulators, partners and effectors of the targets of the principal NPs.

4.2.4. Influence of individual genetic variations

Combinations of sub-potent NPs heavily rely on the synergistic actions of their constituent NPs for improved potencies. Synergistic actions of sub-potent NP combinations typically involve collective modulation of a certain set of the primary targets and the corresponding secondary targets that regulate the primary targets or improve pharmacokinetics of the active NPs. Because of their heavy reliance on the modulation of a corresponding set of secondary targets for achieving sufficiently improved potency, the level of potency improvement of synergistic NP combinations is expected to be sensitively influenced by the genetic variations that alter the expression and activity level of this set of the primary targets and the corresponding secondary targets (150). **Table 4.3** shows the expression profiles of the primary targets and some of the potency-enhancing secondary targets of the selected NP combinations in specific patient groups. The primary targets are expressed in 42%-95% the patients and the secondary targets are expressed in 15%-100% of the patients in different patient groups. Significantly lower percentages of patients in each patient group are expected to have the right set of the primary and the corresponding secondary targets expressed to make them responsive to a particular sub-potent NP combination. Perhaps it is not a coincidence that multi-herb combinations have been frequently prescribed in personalized manner (165, 166) possibly for exploiting certain potency-enhancing modes active in specific patients.

4.3 Summary

This analysis indicates the possibility of synergistically assembling sub-potent NPs into drug-level potent combinations, which can be achieved at low probabilities by the exploration of specific potency-enhancing modes that combine multi-target actions of the principal NPs of sufficient potency (typically within 10-fold range of drug potency levels) against specific disease processes with the enhancement of their bioavailability and/or the modulation of the regulators, effectors and counteractive elements of their targets. The low probabilities for assembling sub-potent NPs into drug-level potent combinations may arise from the difficulties in finding the right combination of NPs with sufficient potency and the appropriate and complementary potency-enhancing MIPs. Moreover, synergistic actions typically involve interactions with multiple sites, targets and pathways which are sensitively influenced by genetic(167), environmental(16), behavioral(168), and scheduling(169) profiles. NP combinations and related therapeutics may be better designed, applied and studied in personalized and environment-dependent manners (170, 171). The efforts in the exploration of NP combinations can be facilitated by expanded knowledge in the activities of NPs (172), MIPs of NPs (73), disease regulations, and potency-enhancing molecular modes that synergistically target key positive (173) and negative (17) regulatory nodes of therapeutic efficacies, and collectively modulate anti-targets and counter-targets (4), compensatory and neutralizing actions (20, 174), and transporter and enzyme mediated pharmacokinetic activities (175).

Table 4. 3 Expression profiles of the primary targets and some of the potency-enhancing secondary targets of the selected natural product combinations in specific patient groups

Natural Product Combination	Target Type	Target	Target Expression Profile in Specific Patient Groups
Tetraarsenic tetrasulfide, Indirubin, and Tanshinone IIA	Primary target of the principal ingredient	PML-RAR	Present in 95% of APL patients (12506013)
	Secondary target for enhancing the potency of the principal ingredient	STAT3	Aberrantly activated in some APL patients (11929748), activated in 71% of AML patients (9679986)
Theaflavin, Theaflavin-3-monogallate, Theaflavin-3'-monogallate, and Theaflavin-3,3' digallate	Primary target of the principal ingredient	JNK	Expressed in 100% of patients with chronic obstructive pulmonary disease (20699612), pJNK expressed in 100% of multiple trauma patients (22677613)
		P38	Expressed in 82% patients with sepsis-induced acute lung injury (17581740), pP38 expressed in 38% of multiple trauma patients (22677613)
	Secondary target involved in the alternative signaling that substitute the targeted pathway of the principal ingredient	Cox2	Expressed in 100% of HBV (15218507) and 100% of HCV (17845691) patients, elevated in 100% of patients with HCV-induced chronic liver disease (18092051)
		ERK	pERK expressed in 15% of colorectal carcinoma (17149612), 39% of mucoepidermoid carcinomas (12937136), 70% of breast cancer (15928662), 79% of mucoepidermoid carcinoma (20664595) patients
Wedelolactone, indole-3-carboxylaldehyde, luteolin, apigenin	Primary target of the principal ingredient	AR	Expressed in 59% of prostate cancer (22500161), 56%-63% of breast cancer (18946753, 22471922), 80% of benign urothelium (22221549), 50% of benign stroma (22221549), 42%-71% of bladder cancer (22221549) patients
	Secondary target for enhancing the potency of the principal ingredient	c-Src	Expressed in 55% of metastatic breast cancer (22716210), 74% of bladder cancer (22353809) , 28% of hormone refractory prostate cancer patients (19447874)
		FGF1R	Expressed in 69%-74% of prostate cancer (17607666), 99%-100% of breast cancer

			(9865904, 9756721) patients
		topoisomerase II	Highly expressed and amplified in 50% and 5%-7% of breast cancer (22240029, 22555090), 31% and 26% of advanced prostate cancer (17363613), 20% and 1.5% of bladder cancer (11304849, 14566826) patients
		CK2	Expressed in the bone marrow of 28% of the patients with transitional cell carcinoma (17977715)
		EGFR	Expressed in 41% of prostate cancer (22500161), 25% of breast cancer (22562124), 33% of triple negative breast cancer (22481575), 66%-96% of bladder cancer (16685269, 19171060) patients
		HER2	Expressed in 1.5%-24% of prostate cancer (19207111, 22500161), 8%-31% breast cancer (10550311, 11344480, 22562124), 62%-98% of bladder cancer (15839918, 16685269) patients
		NF-kB	Expressed in 53% of prostate cancer (21156016), 79% of bladder urothelial carcinoma (18188593), active NF-kB present in 4.4%-43% of breast cancer (16740744) patients
		AkT	pAkT expressed in 45% prostate cancer (19389013) and 33% breast cancer (16464571), highly expressed in 2.6%-14.3% of patients with urothelial carcinoma of the urinary bladder (21707707)
		P53	Expressed in 22%-28% breast cancer (11344480), Overexpressed in 36% of bladder cancer (19171060) patients

Chapter 5: Personalized targeted therapeutics driven by biomarkers

Therapeutic Target Database has a significantly higher number (1,755) of literature-reported biomarkers covering more variety of disease conditions (365) than those in the existing biomarker databases (89-91) and thus complements those databases that primarily include molecular biomarkers of specific disease classes such as the infectious disease biomarker database (IDBD) (89, 91) or clinically prioritized sets (90). The more extensive coverage of potential biomarkers and the convenient access through ICD codes make TTD a useful tool to analyze the biomarker information.

For personalized treatment using targeted therapeutics, the stratification of patients plays an essential role. In order to classify the patients according to their responses to treatment, biomarker information can be utilized. Chapter 5.1 is devoted to the analysis of biomarker information based on which a more refined disease classification system will be suggested.

Furthermore, in chapter 5.2, based on the evaluation of non-invasive biomarkers, their possible application in mobile health (mhealth) technologies is also proposed for delivering healthcare at reduced costs and for facilitating more precise and personalized therapeutics.

5.1 More refined classification of patient subpopulations for personalized targeted therapeutics

Biomarkers have been developed as non-invasive tests for early detection and indication of disease risks, monitoring of disease progression and recurrence, and classification of disease subtypes and patient subpopulations for providing the most appropriate treatments (176-178). As many therapies have been found to elicit markedly different clinical responses in individual patients (179, 180), there is a particular need for more biomarkers capable of predicting drug response in individual patients, which has led to intensive efforts in the discovery of such biomarkers (27, 181). **Table 5.1** gives examples of the approved and clinically tested biomarkers for facilitating the prescription of a particular drug to specific patient subpopulation.

Table 5. 1 Approved and clinically tested biomarkers for facilitating the prescription of a particular drug to specific patient subpopulation

Disease	Therapeutic target	Biomarker for the targeted therapeutics	Patient subpopulation likely responsive to targeted therapeutics	Drug therapy specific for patient subpopulation
Acute promyelocytic leukemia (APL)	PML–RAR	PML–RAR (gene translocation)	APL with PML–RAR α t(15:17) translocation	<i>All trans</i> retinoic acid
Alzheimer's	PPAR	apolipoprotein E and TOMM40 genotypes and age.	Mild cognitive impairment due to Alzheimer's disease	
Breast cancer	HER2	HER2 (gene amplification)	HER2 amplified and/or over-expressed breast cancer	Trastuzumab
	Estrogen receptor	Estrogen receptor (protein expression)	ER overexpressed breast cancer	Tamoxifen
	PARP	BRCA1/2 (mutation)	Breast cancer defective in BRCA1 or BRCA2	Olaparib, veliparib
Chronic myeloid	BCR–	BCR–ABL (gene	Philadelphia chromosome	Imatinib,

leukemia (CML)	ABL	translocation)	and absence of BCR–ABL catalytic domain mutation in CML	dasatinib, nilotinib
Colorectal cancer	EGFR	EGFR(mutation, overexpression), KRAS (mutation)	EGFR overexpression and/or mutation, absence of KRAS mutations in colorectal cancer,	Cetuximab, panitumumab
Non-small-cell lung cancer (NSCLC),	EGFR	EGFR (kinase domain mutation)	EGFR mutations, absence of KRAS mutations in NSCLC	Erlotinib, gefitinib
	ALK	ALK (rearrangements)	Rearranged ALK gene in NSCLC	Crizotinib
Melanoma	BRAF	BRAF V600E (mutation)	Melanoma with RAF V600E mutation	Vemurafenib, Dabrafenib
	MEK	BRAF mutations	Melanoma with RAF mutations	Trametinib
Postmenopausal osteoporosis	RANK ligand	Postmenopausal women with persistent total hip, femoral neck, or lumbar spine BMD T-scores –1.8 to –4.0, or clinical fracture (pharmacodynamic biomarker)	Post-menopausal osteoporosis at high risk for fractures	Denosumab

Targeted therapeutics is naturally linked to molecular-based and cell-based disease-classification systems (e.g. trastuzumab for HER2+ breast cancer and imatinib for Ph+ chronic myelogenous leukemia). From the examples of the approved and clinically tested drug response biomarkers in **Table 5.1**, it seems feasible to incorporate target and biomarker codes into the ICD codes for more refined classification of patient subpopulations responsive to a particular targeted therapy. To further explore this feasibility, the ICD codes for various biomarkers in TTD have been analyzed to check if the current disease classification ICD codes are able to differentiate the different subtypes of diseases and if subtype specific biomarkers are available.

The results of analysis are tabulated in **Table 5.2**. Some known molecular and cell-based disease-subtypes have ICD-10 codes but many are un-coded (**Table 5.2**). Most of the common leukemia types, such as chronic lymphocytic leukemia, acute myelogenous leukemia and chronic myelogenous leukemia, generally have their subtypes coded in ICD system, while many other cancer types like breast cancer and lung cancer do not have coded subtypes.

Though some are yet to be clinically-tested, biomarkers are available for the majority of the known molecular-based and cell-based disease-subtypes. Many more are needed for comprehensive coverage of patient sub-populations. For instance, HER2+ breast cancer need be further divided into HER2E-mRNA and luminal-mRNA subgroups based on a 302-gene multi-marker set (182).

Hence, the analysis of current biomarkers in TTD and ICD classifications suggests that biomarker, target and drug information may be incorporated into the ICD codes for coding these subclasses and refining patient and drug-response sub-populations to facilitate the diagnosis, prescription, monitoring and management of personalized medicine.

Table 5. 2 Examples of diseases and their molecular or cell-based subtypes, ICD codes (marked as NA if unavailable), and the availability (A) or unavailability (NA) of the corresponding diagnostic, prognostic and theragnostic biomarkers and if one or more biomarkers are in clinical use or trial

Disease or Disease Type (ICD-10 Code)	Molecular/Cell-Based Subtype (ICD-10 Code)	Diagnostic Biomarkers	Prognostic Biomarkers	Theragnostic Biomarkers
Breast cancer (C50.0-50.9)	Basal-like ER-/PR-/HER2- (NA)	A	A (clinical trial)	A (clinical trial)
	Luminal types ER+ (C50.X + Z17.0)	A (clinical use)	A (clinical use)	A (clinical use)
	Luminal A ER+ and low grade (NA)	A	A	NA
	Luminal B ER+ but often high grade (NA)	A	A	A
	Luminal ER-/PR+ (NA)	A	A (clinical use)	A (clinical use)
	HER2+	A (clinical use)	A (clinical use)	A (clinical use)
	Claudin-low	A	NA	NA
Lung cancer (34.0-34.9)	Non small cell lung carcinoma (NA)	A	A (clinical use)	A (clinical use)
	NSCLC subtype adenocarcinoma (NA)	A	A	A
	NSCLC subtype squamous-cell lung carcinoma (NA)	A	A	NA
	NSCLC subtype large-cell lung carcinoma (NA)	NA	NA	NA
	Small cell lung carcinoma (NA)	A	A (clinical trial)	NA
Acute lymphoblastic leukemia (C91.0)		A (clinical use)	A (clinical use)	A (clinical use)
	Precursor B acute lymphoblastic leukemia (NA)	NA	A	NA
	Precursor T acute lymphoblastic leukemia (NA)	A	NA	NA
	Burkitt's leukemia (C91.A)	A	NA	NA
	Acute biphenotypic leukemia (C95.0)	A	NA	NA
Chronic lymphocytic leukemia (C91.1)		A (clinical use)	A	A
	B-cell prolymphocytic leukemia (C91.3)	A	NA	NA
	T-cell prolymphocytic leukemia (C91.6)	A	NA	NA
Acute myelogenous leukemia (C92.6, C92.A)		A (clinical use)	A (clinical trial)	A (clinical trial)
	Acute promyelocytic leukemia (C92.4)	NA	A (clinical use)	A (clinical use)
	Acute myeloblastic leukemia (C92.0)	A	A	NA
	Acute megakaryoblastic leukemia (C94.2)	A	NA	NA
Chronic myelogenous leukemia (C92.1)		A (clinical use)	A (clinical use)	A (clinical use)
	Chronic monocytic leukemia (C93.1)	NA	A	NA
Large granular lymphocytic leukemia (C91.Z)		A	NA	NA
	T-cell large granular lymphocytic leukemia (NA)	A	NA	NA
	NK cell large granular lymphocytic leukemia (NA)	NA	NA	NA
Other types of leukemia	Adult T-cell leukemia (C91.5)	A	A	NA
	Hairy cell leukemia (C91.4)	A	A	A

However, many of the existing biomarkers are based on the profile of a single gene. For highly heterogeneous diseases such as cancers, single gene biomarkers are highly limited in their coverage of drug escape mechanisms, and multi-markers may be needed for more sufficient coverage of drug escape mechanisms and for more accurate classification of patient subpopulations in stratified and personal medicines. For instance, BRAF^{V600E} inhibitor dabrafenib have shown improved progression-free survival in BRAF^{V600E} metastatic melanoma patients (183) and outperformed MEK inhibitor trametinib (184) due in part to its specificity to BRAF^{V600E} tumors with a greater therapeutic window (185). However, drug resistance still emerges in dabrafenib-treated BRAF^{V600E} metastatic melanoma patients within months (183). These are primarily due to tumor activation of several BRAF inhibitor escape pathways (185-187). Therefore, the use of a single gene biomarker, BRAF^{V600E} mutation, is insufficient for classifying melanoma patient subpopulations responsive to dabrafenib therapy, and multi-markers need to be introduced for adequately covering active drug escape mechanisms in BRAF^{V600E} metastatic melanoma patients.

Apart from the literature-reported biomarkers, the profiles of various known drug resistance mutations (188-190) and drug response regulators (e.g. the genes promoting drug bypass signaling(191, 192) or hindering drug actions(193) have been studied for predicting drug resistance, which may be potentially explored as drug response biomarkers. In particular, the collective profiles of these drug

response regulators may be considered as potential multi-markers for predicting individual patient's response to drug treatment. For instance, a recent study has shown that collective analysis of the mutational, amplification and expression profile of the 16 literature-reported EGFR tyrosine kinase inhibitor bypass pathway regulators outperforms individual profiles in classifying 53 NSCLC cell-lines sensitive or resistant to EGFR tyrosine kinase inhibitors gefitinib, erlotinib, and lapatinib (97).

Therefore, from the analysis of literature, it could be envisioned that multi-markers which cover drug escape mechanisms could potentially be incorporated in the disease classification code for more personalized and stratified targeted therapies.

5.2 Non-invasive biomarker and their applications to healthcare

5.2.1 Background

There have been intensifying efforts to explore mobile health (mhealth) technologies for delivering healthcare at reduced costs and for facilitating more precise and personalized medicine (194-196). These efforts have led to 73 apps endorsed (**Supplementary Table S5.1**) and additional ones reviewed (194) by the US Food and Drug Administration for self-diagnosis of acute diseases and monitoring of chronic conditions (194) based on such physiological biomarkers as body

temperature, pulse, electrocardiography, spirometry, blood pressure, otoscopy and brainwave (197, 198) (199-202)) and such conventional molecular biomarkers as glucose and urine protein contents (203) (200) (202).

Although these physiological and conventional biomarkers cover many disease conditions, their coverage is limited particularly for cancers, infectious, respiratory, digestive, endocrine and nervous system diseases, as indicated by the disease-coverage profiles of the 73 FDA endorsed, 102 physiological and conventional molecular biomarkers described in the literatures (**Figure 5.1**). Additional biomarkers are needed for fulfilling the potential of mhealth technologies.

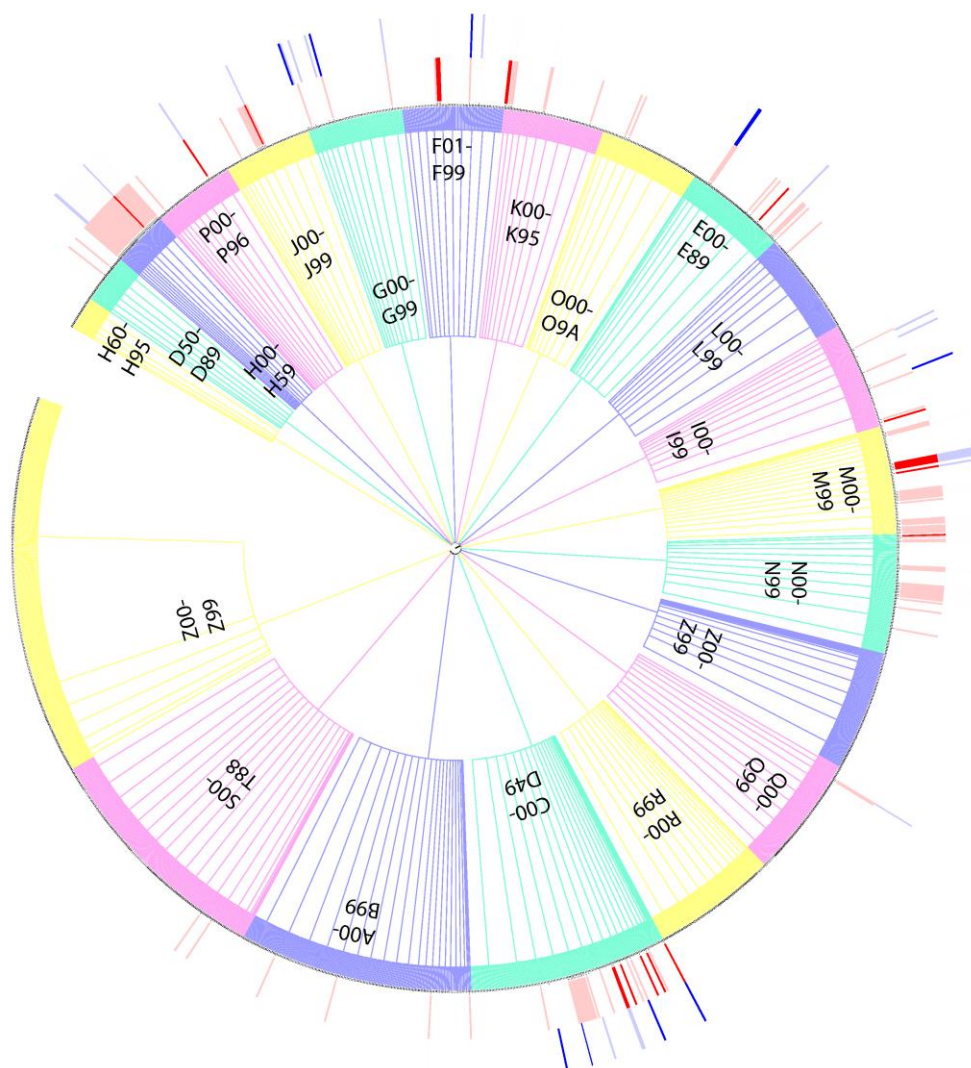


Figure 5. 1 Disease-coverage profiles of the biomarkers.

664 (27 in clinical trial or use) non-invasive molecular biomarkers are colored in light (deep) red. The 94 (13 in clinical trial or use and 73 FDA endorsed apps) physiological and conventional biomarkers are colored in light (deep) blue. Each leaf in the tree represents a specific ICD code. The details of ICD are displayed in Table 2.2.

New genetic, proteomic and metabolomic molecular biomarkers have been discovered and investigated for diagnosing and monitoring diseases, directing treatments and predicting patient responses (178, 204-207). Of immediate relevance to mhealth are the hundreds of literature-reported non-invasive molecular biomarkers from urine, breath, saliva, tear, feces, sputum and oral mucosa samples

collected in TTD, which significantly expand the disease coverage as indicated by the disease-coverage profiles of the 664 (27 clinical trial) non-invasive molecular biomarkers with respect to those of the 866 (73 FDA endorsed) physiological and conventional biomarkers (**Figure 5.1**). Many biomarkers are detectable by the new biomarker-detection technologies that become increasingly portable, fast, user-friendly, inexpensive and accurate (208-213). Efforts have been directed at the exploration of these biomarkers and new technologies for potential mhealth applications (208, 211, 212, 214, 215).

There are questions about whether these biomarkers combined with new technologies are ready for mhealth applications. One is whether the new technologies are sufficiently sensitive, fast and inexpensive for biomarker detection under the typically low sample volume and biomarker concentration conditions. Another is the relevance and accuracies of the literature-reported non-invasive molecular biomarkers for the highly-prevalent disease conditions in need of mhealth tools. The third is how the healthcare providers cope with the increased workload resulting from widespread use of mhealth devices.

These questions can be probed by analysing the literature-reported biomarker detection capability (detection sensitivity, required sample volume, test time, and cost) of the new technologies combined with cellphone or the equivalent imaging devices, and the relevance (disease coverage and patient populations) and accuracies

of the literature-reported non-invasive molecular biomarkers for mhealth-based disease detection and monitoring. The feasibility and potential issues of workload reduction by developing and using a digitally-coded biomarker, disease and therapeutic information processing system for electronic pre-screening of the mhealth biomarker readings will also be discussed.

5.2.2 Evaluation of new biomarker-detection technologies

The new biomarker-detection technologies combined with cellphone or the equivalent imaging devices have been explored for detecting at least 14 molecular biomarkers including 5 non-invasive ones (**Table 5.3**), 50% of which can be detected at low concentrations (0.3-60 pg/mL and 10-20ng/mL for 4 and 3 biomarkers respectively). Although these detectable concentrations are roughly 10-fold higher than those of the conventional technologies (210), at least two are below the corresponding thresholds for non-invasive detection (210, 216). For the biomarkers with higher detectable concentrations, at least one is below the corresponding threshold for non-invasive detection (212). The detection of 64.3% of the biomarkers requires significantly lower sample volumes (0.5-12uL) and shorter detection time (10-60 min) than the typical volumes (100-300uL)(212, 217) and detection times (up to 4h) (210) of the conventional technologies. The cost for the relevant biomarker detection devices is in the range of \$300-\$600 US dollars ((218)).

Therefore, the new technologies are fairly sensitive, efficient, and inexpensive for detecting some of the non-invasive biomarkers in potential mhealth applications.

5.2.3 The relevance and accuracies of the non-invasive molecular biomarkers for mhealth applications

Analysis of the 664 literature-reported non-invasive molecular biomarker (**Table 5.4**) showed that 546 and 183 biomarkers are for the diagnosis and prognosis of 85 and 45 disease conditions respectively. In particular, 31 (36.5%) and 14 (31.1%) disease conditions are covered by higher number (4-22) of biomarkers and 10 (11.8%) and 6 (13.3%) disease conditions by clinically-validated (3 and 1 clinically-used, and 7 and 5 in clinical trial) biomarkers. Many of these disease conditions affect large populations worldwide.

Specifically, 2 out of 2 (100%), 3 out of 10 (33.3%) and 4 out of 15 (26.7%) acute disease conditions and 5 out of 11 (45.5%), 12 out of 22 (54.5%) and 12 out of 44 (27.3%) chronic conditions with clinically-validated, higher and lower number of biomarkers respectively are common diseases, affecting more than 1 million US people or having more than 200,000 new incidences each year in US.

Therefore, exploration of these biomarkers in mhealth applications is expected to have a significant impact on improving the efficiency and quality of the management of these disease conditions.

The accuracies of the 88 (29.7%) of the 296 diagnostic biomarkers in diagnosing 43 disease conditions, and those of the 24 (25.5%) of the 94 prognostic biomarkers in prognosing 14 disease conditions have been reported (**Table 5.4**). The reported biomarker sensitivities (the likelihood for detecting diseases) and specificities (the probability for correctly screening negatives) (the probability that a positive signal is correct) is $\geq 75\%$ (majority $\geq 85\%$) and $\geq 62.5\%$ (majority $\geq 80\%$), and the reported prognosis sensitivities and specificities are $\geq 80\%$ and $\geq 62.8\%$ (majority $\geq 80\%$) respectively. Hence, the accuracies of the majority of these biomarkers are at or close to the $\geq 90\%$ sensitivity and $\geq 90\%$ specificity levels required for the good biomarkers (219). In comparison, the sensitivity and specificity of conventional screening methods used in clinic are in some cases even lower than the reported biomarkers (**Table 5.5**). Hence, these biomarkers may be potentially useful as pre-screening tools for identifying potential patients in need of further attention and test.

5.2.4 A digitally-coded biomarker, disease and therapeutic information processing system

There are concerns about the increased workload arising from widespread use of mhealth devices (1). However, mhealth devices as digital tools may conveniently facilitate electronic pre-screening of the biomarker readings for identifying potential patients likely in need of further attention of the healthcare providers, which helps to significantly reduce the workload. A digitally-coded biomarker, disease and therapeutic information processing system may be developed for automatically receiving, processing and pre-screening the biomarker readings transmitted from mhealth devices, and, upon detecting alert signals, automatically informing healthcare providers for further evaluations and actions.

It is feasible to develop such a system because some of the needed basic tools are in place. These include the International Classification of Diseases (ICD) codes for defining, studying and managing diseases and treatments,(93) the Systematized nomenclature of medicine (SNOMED) for clinical documentation and reporting(220), the Unified medical language system (UMLS) for biomedical terminology(221), the Therapeutic target database (TTD) biomarker and target information and links to the ICD and drug codes (222), and the Drugbank drug information (40).

5.2.5 Future work

Molecular biomarker based mobile health technologies have the potential to significantly improve the efficiency and quality of healthcare for diverse range of disease conditions that cannot be solely covered by physiological and conventional molecular biomarkers. Some of these biomarkers are fairly accurate, sensitive and relevant for mhealth applications. The new technologies enable the exploration of these biomarkers for mhealth applications and the development of electronic systems for efficient management of mhealth activities.

Further efforts are needed for additional information refinement and integration, and the determination and clinical validation of biomarker thresholds for pre-screening purposes. Other obstacles include the potential complications in following the testing protocols, and the possible issues arising from the missed recognition or misrecognition of disease conditions by an electronic system, lack of data security and lack of proper regulation standards.

Table 5. 3 New biomarker-detection technologies.

Biomarker	Biomarker Source	Disease condition	Biomarker Detection Technology	Product Cost	Use of Phone	Detection limit	Minimum Sample Volume	Detection Time	Reference
interferon-gamma	N/A*	latent tuberculosis	An opto-acoustic immunoassay + mobile phone technologies (a surface acoustic wave transducer, a CMOS camera, a LED)	low cost	Y*	1 pM	N/A	10 min	21725557
Bacterial DNA	N/A	bacterial infection	A disposable microfluidic chip with primers + a fluorescence detector + smartphone	\$350-\$600	Y	760 DNA copies per uL	30 uL	30 min	22374412
N-terminal proBNP molecule	blood	heart failure	A disposable biomarker sensing element + HDR image acquisition technique	N/A	Y	60 pg/mL	150 uL	12 min	20926279
PfHRP2	blood	malaria	A disposable microfluidic chip + smartphone with embedded circuit	N/A	Y	16 ng/mL	0.5 uL	15 min	23689554
lactoferrin	tear	disorders of the	An inkjet-printed micro	\$1 per testing	M*	0.3 mg/mL	2.5 uL	15 min	24482793

		corneal epithelium	fluidic paper-based analytical device + digital camera	sheet + cost of digital camera					
HE4	urine	ovarian cancer	Paper-based ELISA + smartphone	N/A	Y	19.5 ng/mL	100 uL	5 h (can be reduced to 15 min)	21881677
VEGF	inner eye humor	ophthalmologically relevant diseases	Paper-based ELISA + scanner	cost of paper-ELISA + \$100 for scanner	M	33.7 fg/mL	2 uL	44 min	24484673
HIV-1 envelope antigen gp41	N/A	HIV infection	Paper-based ELISA + scanner	cost of paper-ELISA + \$100 for scanner	M	18 pM/mL	12 uL	51 min	20512830
adenovirus DNA	N/A	Viral infection	A microfluidic capillary array + an optical signal amplifier (multi-wavelength LEDs) + smartphone	\$180 for capillary array + cost of LED and smartphone	Y	0.4-5 ug/mL	10 uL	N/A	23928092
anti-Leishmania antibodies	canine blood	leishmaniasis	Paper-based ELISA + scanner	cost of paper-ELISA + scanner	M	1 mg/mL	uL range	60 min	24521980

Mycobacterium tuberculosis nucleic acids	N/A	tuberculosis	Paper-based Au-nanoprobes + smartphone	N/A	Y	N/A	N/A	2 h 30 min	24521980
MMP9	urine	colorectal cancer	Paper lateral flow assay + smartphone/scanner	\$2.60 + cost of cellphone	Y	1 mg/mL	5 uL	N/A	24567404
thrombin	urine	thrombosis	Paper lateral flow assay + smartphone/scanner	\$2.60 + cost of cellphone	Y	1 mg/mL	5 uL	N/A	24567404
apolipoprotein A1	urine	bladder cancer	Magnetic bead-based ELISA microfluidic chip	lower costs than conventional ELISA	M	10 ng/mL	14.5 uL	40 min	24484673

*N/A: Not available, unspecified. Y: Smartphones are used. M: Smartphones could be used alternatively.

Table 5. 4 Diseases covered by non-invasive molecular biomarkers

Disease or Disease Type	ICD	diag/ prog	Molecular Type (No of Biomarkers)	Sources	New Tech Feasibility	Highest Sensitivity	Highest Specificity	AUC	Prevalence	Acute/ Chronic	Rare/ Common
Pulmonary tuberculosis	A15.0	diag	P (1)	Sa		81.8%	81.4%		P:World(8.6 M),USA(9,945),UK(0.5 M)	c	Ra
Sepsis	A41.9	diag	P (3)	U	ELISA				P:USA(660,000)	a/c	
Acute hepatitis E	B17.2	diag	P (8)	U	ELISA			0.89	I:World(3 M)	a	Ra
HIV infection	B20	prog	P (6)	U	ELISA	94.0%	71.0%		P:World(35.3 M),USA(1.15 M),UK(2.2 M)	a/c	Co
HIV infection	B20	ther	P (6)	U	ELISA	94.0%	71.0%		P:World(35.3 M),USA(1.15 M),UK(2.2 M)	a/c	Co
Kala-azar	B55.0	diag	P (1)	U	ELISA				I:World(0.5 M)	c	Ra
Upper gastrointestinal cancer	C15-C26	diag	P (4)	U		86.0%	80.0%		I:USA(Esophageal Cancer 17,990)	c	Ra
Gastric cancer	C16	diag	P (1)	U		79.0%	100.0%	0.97	I:World(951,594),USA(21,155),UK(139,667)	c	Ra
Colorectal cancer	C18-C21	diag	P (1), Sm (1)	F, U	ELISA	73.0-83.0%	82.0%		I:World(1360602),USA(134,349),UK(447,136)	c	
Hepatocellular Carcinoma	C22.0	diag	P (2), Sm (1)	U		61.0%	92.0%		I:World(782,451),USA(30,449),UK(63462)	c	Ra
Cholangiocarcinoma	C22.1	diag	P (1)	U		83.0%	79.0%	0.87	I:USA(1.67 in 100,000)	c	Ra

Lung cancer	C33-C34	diag	Sm (2)	Br		84.5%	80.0%		P:USA(214,226)	c	
Lung cancer NSCLC type	C33-C34	prog	Sm (1, CT)	U					P:USA(214,226)	c	
Oral squamous cell carcinoma	C44.02	diag	Sm (2), P(2)	Sa		92.3%	91.7%		I:World(640,000),USA(54,000)	c	
Oral squamous cell carcinoma	C44.02	prog	P (1)	Sa, Sk					I:World(640,000),USA(54,000)	c	
Breast cancer	C50	prog	Pep (1)	U	ELISA				I:World(1676633),USA(232,714),UK(464,202)	c	Co
Ovarian cancer	C56	diag	Sm (1)	U	ELISA	70.0%	75.0%		I:World(238,719),USA(20,874),UK(65,584)	c	Ra
Prostate cancer	C61	diag	Sm (1, CT), Pep (1, CT)	U	ELISA				I:World(1111689),USA(233,159),UK(417,137)	c	Co
Renal cell carcinoma	C64	diag	P (12)	U	ELISA	100.0%	100.0%	1	I:World(245,000),USA(65,000),UK(91,000)	c	
Kidney cancer	C64.9	diag	P (2, CT 2)	U	ELISA	100.0%	100.0%	1	I:World(337860),USA(58,222),UK(115,252)	c	
Bladder cancer	C67	diag	P (1)	U	ELISA	85.7%			I:World(429,793),USA(68,639),UK(151,297)	c	
Bladder cancer	C67	prog	P (4)	U	ELISA				I:World(429,793),USA(68,639),UK(151,297)	c	
Malignant primary brain	C71	diag	Sm (1, CT)	U	ELISA				I:World(256,000),USA(69,720),UK(57,100)	c	

tumor											
Carcinoid tumor	C75, E34.0	diag	Sm (1)	U		35.0%	88.0%		I:World(12,000)	c	Ra
Bone metastases	C79.51	prog	P (1, CT), Pep (1, CT), Sm (1, CT)	U	ELISA					c	
Multiple myeloma	C90.0	diag	P (1)	U	ELISA	88.9%	83.3%		I:World(114,000),USA(24,050),UK(38,900)	c	Ra
Multiple myeloma	C90.0	prog	Pep (1)	U	ELISA				I:World(114,000),USA(24,050),UK(38,900)	c	Ra
Henoch-Schonlein purpura	D69.0	prog	P (1)	U	ELISA				I:World(10-22 in 100,000)	a	
Acute graft-versus-host disease	D89.8	diag	P (9)	Sa, U, Sk	ELISA				I:World(5500)	a	Ra
Type 1 diabetes	E10	diag	P (2)	U	ELISA				P:World(11-22 M),USA(3 M),UK(112,000)	c	Co
Type 1 diabetes	E10	prog	P (1, combi 4)	U	ELISA			0.89	P:World(11-22 M),USA(3 M),UK(112,000)	c	Co
Diabetes	E10, E11	diag	P (2, combi 261)	U		~91%	~78%	0.94		c	Co
Diabetic Nephropathy	E10.2, E11.2, E12.2,	diag	P (7)	U	ELISA	81.4%	62.5%		P:World(20% - 40% of diabetes)	c	Co

	E13.2, E14.2										
Diabetic Nephropathy	E10.2, E11.2, E12.2, E13.2, E14.2	prog	P (3)	U	ELISA				P:World(20% - 40% of diabetes)	c	Co
Diabetes mellitus type 2	E11	diag	P (10)	U	ELISA				P:World(90 % diabetes)	c	Co
Diabetes mellitus type 2	E11	prog	P (3)	U	ELISA				P:World(90 % diabetes)	c	Co
Diabetes insipidus	E23.2	diag	P (1)	U					I:World(3 per 100,000)	a/c	
Aldosteronism	E26.02	diag	P (1)	U					P:,USA(<200,000)	c	Ra
Mucopolysacchari doses	E76	diag	Sm (2)	U	ELISA				P:USA(200)	c	Ra
Mucopolysacchari doses	E76	diag	Sm (2)	U	ELISA				P:USA(200)	c	Ra
Idiopathic hypercalciuria(IH)	E83.52	diag	P (1)	U	ELISA					c	
Cystic fibrosis	E84	diag	Sm (2)	Br		93.8%	69.2%		P:World(70,000),USA(30,000)	c	Ra
Cystic fibrosis	E84	prog	Sm (3)	Br	ELISA				P:World(70,000),USA(30,000)	c	Ra
Renal light-chain	E85.8	diag	P (1)	U		81.3%	98.0%		I:,USA(1200-3200)	c	Ra

amyloidosis (AL)											
Metabolic syndrome	E88.81	prog	P (1)	U					P:;USA(22.9% of population)	c	Co
Chronic stress	F40-F42	diag	P (1, CT)	U	ELISA	100.0%			P:World(40 M)	c	Co
Enuresis	F98.0	prog	P (1)	U					P:;USA(4-4.5% of children)	c	Co
Parkinson"s disease	G20	prog	Sm (1)	U	ELISA				P:World(10 M),USA(1 M),UK(6.7 M)	c	Co
Multiple sclerosis	G35	diag	P (1)	U	ELISA				P:World(30 in 100 000),USA(400,000),UK(80 in 100 000)	a/c*	
Obstructive sleep apnea syndrome	G47.33	diag	Sm (1)	Sa					P:World(3%-7%),USA(4% in men, 2% in women)	c	Co
Encephalopathy	G93.4	diag	Sm (1), P (1)	U	ELISA	99.0%	97.0%			a	
Encephalopathy	G93.4	prog	Sm (1), P (1)	U	ELISA					a	
Ocular allergy	H00-H59	diag	Pep (2), P (3)	T	ELISA				P:World(2 M)	a*/c	Ra
Ocular allergy	H00-H59	prog	P (1)	T	ELISA				P:World(2 M)	a*/c	Ra
Dry eye disease	H16.229	diag	Pep (2, CT 1), P (19)	T	ELISA	85.0%	94.0%		P:World(4.88 M)	c	
Dry eye disease	H16.229	prog	Pep (2), P (1)	T	ELISA				P:World(4.88 M)	c	
Glaucoma	H40-H4	diag	P (1)	Eye					P:World(60 M),USA(2.2 M)	a/c*	Co

	2										
Chronic renovascular hypertension	I15.0	diag	P (1)	U	ELISA					c	
Pulmonary arterial hypertension	I27.0, I27.2	ther	Sm (1)	Br					P:,USA(260,000)	a*/c	
Atrial fibrillation	I48	diag	Sm (1)	U					P:World(33.5 M),USA(2.66 M)	a/c	Co
Heart failure	I50	diag	P (1, CU), Sm (1)	Hair, U					P:World(26 M),USA(5.1 M),UK(3.5 M)	a/c	Co
Heart failure	I50	prog	P (2)	U					P:World(26 M),USA(5.1 M),UK(3.5 M)	a/c	Co
Kidney function decline with atherosclerosis	I75.81	diag	P (1)	U						c	
Deep vein thrombosis(DVT) and pulmonary embolism(PE)	I82.4,I82.5	diag	P (1)	U	ELISA	100.0%	85.0%	0.97	P:USA(300,000-600,000)	c	
Chronic obstructive pulmonary disease	J40-J44, J47	diag	Sm (3)	Br, U					P:World(64 M),USA(12.7 M),UK(1.5-10% of population)	c	Co

Chronic obstructive pulmonary disease	J40-J44, J47	prog	Sm (2)	Br					P:World(64 M),USA(12.7 M),UK(1.5-10% of population)	c	Co
Asthma	J45	diag	Sm (4), P (1), Cell (2)	Br, Sp	ELISA	73.6-86.0%	88.0%		P:World(235 M),USA(25 M),UK(30 M)	c	Co
Asthma	J45	prog	Sm (2), P (1) Sm+P (1, CT), Cell (1), Sm+Cell (1)	Br, Sp	ELISA				P:World(235 M),USA(25 M),UK(30 M)	c	Co
Fibrosing alveolitis	J84.1	prog	Sm (1)	Br					P:USA(14-27.9 in 100,000),UK(1.25-23.4 per 100,000)	c*/a	Ra
Dental caries	K02	diag	P (1), Pep (1)	Sa					P:World(23.7% adult),USA(15.6% children),UK(59% population)	c	Co
Acute appendicitis	K35-K37	diag	Proten (9)	U		95.0%	100.0%	0.98	I:USA(680,000)	a*/c	Co
Acute appendicitis	K35-K37	prog	P (2)	U	ELISA	~82%	~68%	0.8	I:USA(680,000)	a*/c	Co
Crohn's disease	K50	prog	P (2)	U					P:World(0.1-16 in 100,000)	c	Ra
Inflammatory Bowel Disease	K50,K51	diag	P (12, CU 2), Sm (1)	Br, F	ELISA	80-98%, 94%	82-96%, 76%		P:World(0.396% population),USA(1.4	c	Co

									M),UK(2.5-3 M)		
Inflammatory Bowel Disease	K50,K51	prog	P (16, CU 2)	F	ELISA	80-90%, 70-100%	82-83%, 44-100%		P:World(0.396% population),USA(1.4 M),UK(2.5-3 M)	c	Co
Inflammatory Bowel Disease	K50,K51	ther	P (2)	F	ELISA				P:World(0.396% population),USA(1.4 M),UK(2.5-3 M)	c	Co
Acute pancreatitis	K85	diag	P (8)	U	ELISA	100.0%	96.0%		I:USA(32-44 in 100,000)	a	
Acute pancreatitis	K85	prog	P (11)	U	ELISA	91.7%	89.7%	0.81	I:USA(32-44 in 100,000)	a	
Pancreatitis	K85, K86.0-K86.1	diag	P (2)	U		81.0%	97.0%		I:USA(13-45 acute + 5-12 xhronic in 100,000)	a/c	
Psoriasis	L40	diag	P (2), miR (4), cell (1)	Sk	ELISA				P:World(125 M),USA(7.5 M),UK(11 M)	c	Co
Arthritis	M00-M25	diag	P (17)	U		~85%	~100%	1	P:World(1% of population),USA(52.5 M)	c	Co
Arthritis	M00-M25	prog	P (1)	U	ELISA				P:World(1% of population),USA(52.5 M)	c	Co
Osteoarthritis	M15-M19, M47	diag	P (3)	U	ELISA	74.6%	85.7%	0.84		c	
Osteoarthritis	M15-M19,M47	diag	Sm (1), Pep (1), Modified Pep (2, CT 1)	U					P:World(26.9 M)	c	Co

Osteoarthritis	M15-M19,M47	prog	Sm (1), Pep (3), Modified Pep (2)	U					P:World(26.9 M)	c	Co
Knee osteoarthritis	M17	diag	P (1)	U	ELISA					c	
Kawasaki disease	M30.3	diag	P (14)	U	ELISA	~92%	~95%	0.98	P:USA(9-19 in 100,000 children ?5 years)	a	Ra
Systemic lupus erytematosus	M32	diag	P (2)	U	ELISA				P:USA(161,000-322,000)	c	
Systemic lupus erytematosus	M32	prog	P (3)	U	ELISA	~70%	~89%	0.76	P:USA(161,000-322,000)	c	
Lupus nephritis	M32.1, N08.5	diag	P (3)	U	ELISA	88.5%	46.3%	0.73		c	
Lupus nephritis	M32.1, N08.5	prog	P (5, combi 3), miR (2)	U	ELISA	100.0%	81.0%	0.92		c	
Focal segmental glomerulosclerosis (FSGS)	N00.1, N01.1, N02.1, N03.1, N04.1, N05.1, N06.1, N07.1	diag	P (1)	U	ELISA				P:USA(70,000)	c	Ra
Crescentic	N00.7,	diag	P (1)	U	ELISA	91.7%	90.2%			a	

Glomerulonephritis(GN)	N01.7, N02.7, N03.7, N04.7, N05.7, N06.7, N07.7										
Membranous nephropathy	N02.2	diag	P (1)	U	ELISA	86.0%			I:USA(2000)	c	Ra
IgA nephritis	N02.8	diag	P (65)	U	ELISA	81.7%	73.4%		P:USA(1 in 100,000)	c	Ra
IgA nephritis	N02.8	prog	P (9)	U	ELISA	100.0%	100.0%	1	P:USA(1 in 100,000)	c	Ra
IgA nephritis	N02.8	ther	P (8)	U	ELISA				P:USA(1 in 100,000)	c	Ra
Chronic glomerulonephritis	N03.2	prog	P (1)	U	ELISA	87.5%	90.5%	0.95		c	
Nephrotic syndrome	N04	diag	P (6)	U					P:USA(15 in 100,000 children)	a	Ra
Nephrotic syndrome	N04	prog	P (1)	U					P:USA(15 in 100,000 children)	a	Ra
Nephrotic syndrome	N04	ther	P (1)	U					P:USA(15 in 100,000 children)	a	Ra
Minimal change nephropathy	N04.0	diag	P (1)	U	ELISA				P:USA(1.5-2.3 per 100,000)	a	Ra
Idiopathic	N04.9	diag	P (1)	U						a	

nephrotic syndrome (INS)											
Idiopathic nephrotic syndrome (INS)	N04.9	prog	P (1)	U						a	
Vesicoureteral Reflux	N13.7	diag	P (2)	U	ELISA	67.0%	85.0%	0.77	P:World(1%-2% of children)	c	
Vesicoureteral Reflux	N13.7	prog	P (2)	U	ELISA	81.2%	85.0%	0.88	P:World(1%-2% of children)	c	
Contrast-induced nephropathy	N14.1	diag	P (18)	U	ELISA	73.0%	100.0%	0.92	P:World(<2% of population)	a	
Contrast-induced nephropathy	N14.1	prog	P (2)	U	ELISA	80.0%	75.0%		P:World(<2% of population)	a	
Balkan endemic nephropathy	N15.0	diag	P (4)	U	ELISA	72.3%	84.4%	0.83	P:World(0.5-4.4% of population),USA(<200,000)	c	Ra
Balkan endemic nephropathy	N15.0	prog	P (1)	U	ELISA				P:World(0.5-4.4% of population),USA(<200,000)	c	Ra
Acute kidney injury	N17	diag	P (15, CU 2, CT 3)	U	ELISA	69-100%, 73-100%	85-98%		P:USA(1-7.1% of all hospital admissions)	a	Co
Acute kidney injury	N17	prog	P (2, CT 1)	U	ELISA	>90%	>90%		P:USA(1-7.1% of all hospital admissions)	a	Co
Chronic kidney disease	N18.9	diag	P (2)	U	ELISA				P:World(8-16% of population),USA(20 million)	c	Co
Chronic kidney	N18.9	prog	P (18)	U	ELISA				P:World(8-16% of	c	Co

disease									population),USA(20 million)		
Kidney calculi	N20.0	diag	P (20)	U					P:USA(1 in 11)	c	Co
Urolithiasis	N21.0- N21.9	diag	P (3)	U	ELISA	90.0%	68.0%		P:USA(7% of women and 12% of men)	c	Co
Urolithiasis	N21.0- N21.9	prog	P (1)	U	ELISA				P:USA(7% of women and 12% of men)	c	Co
Interstitial cystitis	N30.10, N30.11	diag	P (7), Sm (2)	U	ELISA	70.0%	72.4%		P:USA(8 million women)	c	Co
Overactive bladder	N32.81	diag	Sm (1), P (4)	U	ELISA				P:World(33 M),USA(22 M)	c	Co
Overactive bladder	N32.81	prog	Sm (1), P (4)	U	ELISA				P:World(33 M),USA(22 M)	c	Co
Urinary tract infection	N39.0	diag	P (1)	U	ELISA				P:USA(1 in 5 women)	a*/c	Co
Dents disease	N39.8	diag	P (66)	U					P:World(250)	c	Ra
Endometriosis	N80	diag	P (1)	U					P:World(6-10% of women)	c	Co
Pre-eclampsia	O11,O1 4	diag	P (9)	U	ELISA				P:USA(3-4% baby-delivery women)	a	Co
Pre-eclampsia	O11,O1 4	prog	P (4)	U	ELISA	~56%	~73%	0.64	P:USA(3-4% baby-delivery women)	a	Co
Bronchopulmonar y dysplasia	P27.1	diag	Sm (3), Pep (1, CT)	Br, U	ELISA	50-85%	61.1-90.0%		I:World(12,000)	c	Ra
Necrotizing enterocolitis	P77	diag	P (3)	U	ELISA				P:World(1-3 in 1,000 infants)	a	Ra

Necrotizing enterocolitis	P77	prog	P (1)	U	ELISA				P:World(1-3 in 1,000 infants)	a	Ra
Primary ciliary dysSesia	Q34.8	diag	Sm (1)	Br					P:USA(25000)	c	Ra
Autosomal dominant polycystic kidney disease	Q61	diag	Sm (1), P (5)	U	ELISA				P:World(12.5 million),USA(0.6 M)	c	
Congenital hydronephrosis	Q62.0	diag	P (1)	U	ELISA	~85%	~90%	0.86		a/c	
Ureteropelvic junction obstruction	Q62.11	diag	P (36)	U					P:World(0.5-1 in 1000 newborns)	c	Ra
Traumatic brain injury (TBI)	S06	prog	P (1)	U	ELISA	90.0%	62.8%	0.78	P:USA(823.7 in 100,000)	a/c	Co
Rejection of renal transplants	T86.1	diag	P (5), P+Pep (1)	U	ELISA	80-92%, 63-100%	77-83%, 63-98%			a	
Rejection of renal transplants	T86.1	prog	P (1)	U	ELISA	84-87%	95-96%			a	

diag:diagnostic, prog:prognostic, ther, theragnostic

P:Protein, Sm:Small molecule, Pep: Peptide, miR:microRNA, CU:Clinical use, CT:Clinical trial,

combi:combination

Breath: Br, Feces:F, Saliva: Sa, Skin:Sk, Sputum:Sp, Tears: T,Urine:U,

A: acute, C:Chronic

Co:Common, Ra:Rare

Table 5. 5 Conventional test performance

Disease	Biomarker/test	Sensitivity	Specificity	Biomarker type	Molecule type	Location
Cervical Cancer mild dysplasia	Pap smear test	68%	75%	diagnostic		tissue
Cervical Cancer moderate/severe dysplasia	Pap smear test	70% - 80%	95%	diagnostic		tissue
Cervical Cancer moderate dysplasia	Self-collected HPV DNA testing	86.20%	80.70%	diagnostic	DNA	tissue
Cervical Cancer severe dysplasia	Self-collected HPV DNA testing	86.10%	79.50%	diagnostic	DNA	tissue
breast cancer	Mammography			diagnostic		tissue
colorectal cancer	Fecal Occult Blood Test (FOBT)	88.20%	89.70%	diagnostic		stool
advanced neoplasm	Fecal Occult Blood Test (FOBT)	61.50%	91.40%	diagnostic		stool
proximal colon cancer	fecal immunochemical testing (FIT)	58.30%	94.50%	diagnostic		stool
colorectal cancer	Stool DNA testing	51.60%		diagnostic	DNA	stool
lung cancer	low-dose helical computed tomography (LDCT)	40% -95%		diagnostic		body
lung cancer	chest radiography			diagnostic		body
Neuroblastoma	vanillylmandelic acid (VMA) and homovanillic acid(HVA) levels	40% -80%	99.90%	diagnostic	small molecule	urine
ovarian cancer	Transvaginal ultrasonography (or transvaginal sonography)		84.90%	diagnostic		body
ovarian cancer	CA125	20% -57%	95%	diagnostic	protein	serum
prostate cancer	Digital Rectal Exam	16.70%		diagnostic		body
prostatic carcinoma	Transrectal Ultrasound	71%-92%	49%-79%	diagnostic		body
prostate cancer	PSA	71%	91%	prognostic	protein	serum

stomach cancer	pepsinogen levels I and II (PGI and PGII)	84.60%	73.50%	diagnostic	protein	serum
stomach cancer	Barium-meal gastric photofluorography		67%-80%	diagnostic		body
stomach cancer	Gastric endoscopy			diagnostic		body
coronary heart disease	cholesterol	83.30%	96.30%	diagnostic	small molecule	blood
type 2 diabetes	hemoglobin A1c	66%	98%	diagnostic	protein	blood
diabetic retinopathy	glucose	75%-80%	75%-80%	diagnostic	small molecule	plasma
diabetic retinopathy	fasting plasma glucose	75%-80%	75%-80%	diagnostic	small molecule	plasma
diabetic retinopathy	hemoglobin A1c	75%-80%	75%-80%	diagnostic	protein	blood
hypertension	blood pressure	100%	70.40%	diagnostic		body
Phenylketonuria	PKU screening test	100%	51%	diagnostic		blood
Phenylketonuria	PKU screening test	100%	98%	diagnostic		blood
thyroid dysfunction	Thyroid-Stimulating Hormone (TSH)	89-95%	90-96%	diagnostic	small molecule	serum

Chapter 6: Concluding remarks

The modern rational drug discovery process starts with the hypothesis that modulation of certain targets may exert therapeutic value and therapeutics directed at those targets are developed to combat diseases. The identification and validation of target led by the knowledge of the molecular basis of diseases in the early steps of drug discovery paves the way for the drug development directed at the specific targets. In comparison with the traditional drugs, rationally designed drugs directed at specific targets show more promising efficacy profile and fewer toxic side effects.

Although our knowledge of the underlying mechanisms of diseases is increasing, modern drug discovery remains a lengthy and costly process. Ways to enhance its productivity are highly desired. In this big data era, the large and complex collection of various targeted therapeutics data call for efficient data management and analysis methods. In this thesis, efforts to update TTD for store, integrate and retrieve reliable data of therapeutics data and various bioinformatics methods to analyse these data for drug discovery were reported.

In Chapter 2, various aspects regarding the therapeutics data in TTD were presented, such as data collection methods, data sources and ways to access data in TTD. The search tools for using the International Classification of Disease ICD-10-CM and ICD-9-CM codes were added to link and retrieve the target, biomarker and drug

information (currently enabling the search of almost 900 targets, 1800 biomarkers and 6000 drugs related to 900 disease conditions). Information of almost 1800 biomarkers for 300 disease conditions were newly added to the TTD in the latest update. The data contents were significantly expanded to cover >2300 targets (388 successful and 461 clinical trial targets), 20 600 drugs (2003 approved and 3147 clinical trial drugs) and 20 000 multitarget agents against almost 400 target-pairs. The updated TTD database enables more convenient data access and will serve the bench-to-clinic communities better by facilitating the discovery, investigation, application, monitoring and management of targeted therapeutics.

Chapter 3 described several methods to learn from the properties and structures of known drugs and inhibitors for better design of multi-target small molecule drugs. In chapter 3.1, the evaluation of three VS methods showed reasonably good dual-inhibitor yields consistently in all models and the dual-inhibitor yields of the target pairs at all similarity levels were comparable among the three methods. The false hit rate of combi-SVM method was comparable and in some cases better than the false hit rate of other VS tools reported in literature. But the dual inhibitor yields tended to show larger variations at decreasing similarity between the drug-binding domains of the target pairs, suggesting that it was more difficult to produce consistent dual inhibitor yields for lower similarity target pairs. And the selectivity of all three methods against individual-target inhibitors tended to be significantly decreased when similarity level of the target pairs is increased. Based on the

evaluation results, the VS tools were good at predicting dual-inhibitors with reasonably high yields and low false hit rate. But the target selectivity performance of these VS tools needed to be further improved, especially in target pairs with high similarity levels in their binding sites.

The VS tools developed in Chapter 3.1 could identify virtual hits from the large chemical libraries, but from hit to lead and from lead to drugs, more methods were needed to shorten the process and to increase the success rate. In chapter 3.2, a hierarchical clustering method was proposed to cluster known drugs in the target specific chemical space spanned by compounds from large chemical libraries. Preliminary investigation seemed to hint that there could be some drug prolific regions showing privileged drug like structure scaffold and drugs tended to have certain properties in comparison with the inhibitors with similar structure scaffold. This method will be further evaluated on more datasets to generate more reliable rules to predict compounds of good structure scaffold and optimal drug properties that could have higher chance to enter clinical trials and become drugs.

A systematic analysis of natural product combinations was detailed in Chapter 4 to learn from nature in search of novel multi-target mechanisms. Through analysing 124 synergistic natural product combinations, and 122 molecular interaction profiles of the 19 natural product combinations with collective potency enhanced to drug level or by >10-fold, it was found that most of the evaluated natural products and

combinations were sub-potent to drugs. And sub-potent natural products could be assembled into combinations of drug level potency, though at relatively low probabilities. Distinguished multi-target modes that modulating primary targets, their regulators and effectors, and intracellular availability of the active natural products were identified and could shed light to the design of multi-target therapeutics.

To reflect the current shift of drug development focus to more personalised targeted therapeutics, the biomarker information was systematically analyzed in Chapter 5. The analysis of current biomarkers in TTD and ICD classifications suggested that biomarker (especially multi-markers), target and drug information may be incorporated into the ICD codes for coding these subclasses and refining patient and drug-response sub-populations. In addition, evaluation of non-invasive biomarkers in literature suggested that molecular biomarker based mobile health technologies have the potential to significantly improve the efficiency and quality of healthcare for diverse range of disease conditions. Many non-invasive biomarkers are fairly accurate, sensitive and relevant for mhealth applications.

The discovery and application of targeted therapeutics increasingly involve collective efforts from multiple bench-to-clinic communities and are moving more and more towards stratified and personalized medicine. The drug, target, biomarker, and other relevant chemical, biological, pharmaceutical and clinical data need to be

more integrated and easily accessible by the multiple bench-to-clinic communities.

Continuous efforts to develop and improve bioinformatics methods for analyzing targeted therapeutics will greatly facilitate drug discovery.

Bibliography

1. Booth B & Zimmel R (2004) Prospects for productivity. *Nat Rev Drug Discov* 3(5):451-456.
2. Hughes JP, Rees S, Kalindjian SB, & Philpott KL (2011) Principles of early drug discovery. *British journal of pharmacology* 162(6):1239-1249.
3. Lindsay MA (2003) Target discovery. *Nat Rev Drug Discov* 2(10):831-838.
4. Overall CM & Kleinfeld O (2006) Tumour microenvironment - opinion: validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. *Nat Rev Cancer* 6(3):227-239.
5. Vidalin O, Muslmani M, Estienne C, Echchakir H, & Abina AM (2009) In vivo target validation using gene invalidation, RNA interference and protein functional knockout models: it is the time to combine. *Current opinion in pharmacology* 9(5):669-676.
6. Stumpf WE (2007) Memo to the FDA and ICH: appeal for in vivo drug target identification and target pharmacokinetics Recommendations for improved procedures and requirements. *Drug Discov Today* 12(15-16):594-598.
7. Langer T & Hoffmann RD (2001) Virtual screening: an effective tool for lead structure discovery? *Curr Pharm Des* 7(7):509-527.
8. Kenny BA, Bushfield M, Parry-Smith DJ, Fogarty S, & Treherne JM (1998) The application of high-throughput screening to novel lead discovery. *Progress in drug research. Fortschritte der Arzneimittelforschung. Progres des recherches pharmaceutiques* 51:245-269.
9. Keseru GM & Makara GM (2009) The influence of lead discovery strategies on the properties of drug candidates. *Nat Rev Drug Discov* 8(3):203-212.
10. Myers PL (1997) Will combinatorial chemistry deliver real medicines? *Current opinion in biotechnology* 8(6):701-707.
11. Gallop MA, Barrett RW, Dower WJ, Fodor SP, & Gordon EM (1994) Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J Med Chem* 37(9):1233-1251.
12. Gordon EM, Barrett RW, Dower WJ, Fodor SP, & Gallop MA (1994) Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J Med Chem* 37(10):1385-1401.
13. Holford NH, Kimko HC, Monteleone JP, & Peck CC (2000) Simulation of clinical trials. *Annual review of pharmacology and toxicology* 40:209-234.
14. Offermann M (2008) The era of targeted therapies. *American family physician* 77(3):294, 296.
15. de Bono JS & Ashworth A (2010) Translating cancer research into targeted therapeutics. *Nature* 467(7315):543-549.
16. Zimmermann GR, Lehar J, & Keith CT (2007) Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov Today* 12(1-2):34-42.
17. Smalley KS, *et al.* (2006) Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases. *Mol Cancer Ther* 5(5):1136-1144.
18. Pilpel Y, Sudarsanam P, & Church GM (2001) Identifying regulatory networks by

- combinatorial analysis of promoter elements. *Nat Genet* 29(2):153-159.
19. Muller R (2004) Crosstalk of oncogenic and prostanoid signaling pathways. *J Cancer Res Clin Oncol* 130(8):429-444.
20. Sergina NV, *et al.* (2007) Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature* 445(7126):437-441.
21. Christopher M., Overall, & Kleifeld O (2006) Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. *Nature Reviews Cancer* 6:227-239.
22. Force T, Krause DS, & Van Etten RA (2007) Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition. *Nat Rev Cancer* 7(5):332-344.
23. Ma XH, *et al.* (2010) Virtual Screening of Selective Multitarget Kinase Inhibitors by Combinatorial Support Vector Machines. *Mol Pharm.*
24. Zheng CJ, *et al.* (2006) Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev* 58(2):259-279.
25. Overington JP, Al-Lazikani B, & Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5(12):993-996.
26. Rask-Andersen M, Almen MS, & Schioth HB (2011) Trends in the exploitation of novel drug targets. *Nat Rev Drug Discov* 10(8):579-590.
27. La Thangue NB & Kerr DJ (2011) Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nat Rev Clin Oncol* 8(10):587-596.
28. Trusheim MR, *et al.* (2011) Quantifying factors for the success of stratified medicine. *Nat Rev Drug Discov* 10(11):817-833.
29. Volzke H, *et al.* (2013) Personalized cardiovascular medicine: concepts and methodological considerations. *Nat Rev Cardiol* 10(6):308-316.
30. Kneller R (2010) The importance of new companies for drug discovery: origins of a decade of new drugs. *Nat Rev Drug Discov* 9(11):867-882.
31. Bunnage ME (2011) Getting pharmaceutical R&D back on target. *Nat Chem Biol* 7(6):335-339.
32. Chataway J, Fry C, Marjanovic S, & Yaqub O (2012) Public-private collaborations and partnerships in stratified medicine: making sense of new interactions. *N Biotechnol* 29(6):732-740.
33. Maliepaard M, *et al.* (2013) Pharmacogenetics in the evaluation of new drugs: a multiregional regulatory perspective. *Nat Rev Drug Discov* 12(2):103-115.
34. Biomarkers Definitions Working G (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 69(3):89-95.
35. Carden CP, *et al.* (2010) Can molecular biomarker-based patient selection in Phase I trials accelerate anticancer drug development? *Drug Discov Today* 15(3-4):88-97.
36. Engelman JA, *et al.* (2007) MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science* 316(5827):1039-1043.
37. Chen Z, *et al.* (2012) A murine lung cancer co-clinical trial identifies genetic modifiers of therapeutic response. *Nature* 483(7391):613-617.
38. Hall IP (2013) Stratified medicine: drugs meet genetics. *Eur Respir Rev* 22(127):53-57.
39. Fugel HJ, Nuijten M, & Postma M (2012) Stratified medicine and reimbursement issues. *Front Pharmacol* 3:181.
40. Law V, *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids*

- Res 42(Database issue):D1091-1097.
41. Zhu F, *et al.* (2009) What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. *J Pharmacol Exp Ther* 330(1):304-315.
 42. Prado-Prado FJ, *et al.* (2009) Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg Med Chem* 17(2):569-575.
 43. Matzen L, *et al.* (2000) 5-HT reuptake inhibitors with 5-HT(1B/1D) antagonistic activity: a new approach toward efficient antidepressants. *J Med Chem* 43(6):1149-1157.
 44. Dessalew N (2009) QSAR study on dual SET and NET reuptake inhibitors: an insight into the structural requirement for antidepressant activity. *J Enzyme Inhib Med Chem* 24(1):262-271.
 45. Ma XH, *et al.* (2010) In-silico approaches to multi-target drug discovery : computer aided multi-target drug design, multi-target virtual screening. *Pharm Res* 27(5):739-749.
 46. Luan X, *et al.* (2011) Exploration of acridine scaffold as a potentially interesting scaffold for discovering novel multi-target VEGFR-2 and Src kinase inhibitors. *Bioorg Med Chem* 19(11):3312-3319.
 47. Marzaro G, *et al.* (2011) Using the TOPS-MODE approach to fit multi-target QSAR models for tyrosine kinases inhibitors. *Eur J Med Chem* 46(6):2185-2192.
 48. Ashburn TT & Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 3(8):673-683.
 49. Miska D (2003) Biotech's twentieth birthday blues. *Nat Rev Drug Discov* 2(3):231-233.
 50. Prentis RA, Lis Y, & Walker SR (1988) Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964-1985). *British journal of clinical pharmacology* 25(3):387-396.
 51. Wenlock MC, Austin RP, Barton P, Davis AM, & Leeson PD (2003) A comparison of physiochemical property profiles of development and marketed oral drugs. *J Med Chem* 46(7):1250-1256.
 52. 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(1-3):3-26.
 53. Wagener M & van Geerestein VJ (2000) Potential drugs and nondrugs: prediction and identification of important structural features. *Journal of chemical information and computer sciences* 40(2):280-292.
 54. Bemis GW & Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39(15):2887-2893.
 55. Wang J & Hou T (2010) Drug and drug candidate building block analysis. *J Chem Inf Model* 50(1):55-67.
 56. Duarte CD, Barreiro EJ, & Fraga CA (2007) Privileged structures: a useful concept for the rational design of new lead drug candidates. *Mini Rev Med Chem* 7(11):1108-1119.
 57. Koch MA, *et al.* (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci U S A* 102(48):17272-17277.
 58. Kong DX, Jiang YY, & Zhang HY (2010) Marine natural products as sources of novel scaffolds: achievement and concern. *Drug Discov Today* 15(21-22):884-886.
 59. Mallinson J & Collins I (2012) Macrocycles in new drug discovery. *Future Med Chem*

- 4(11):1409-1438.
60. Newman DJ & Cragg GM (2007) Natural products as sources of new drugs over the last 25 years. *J Nat Prod* 70(3):461-477.
61. Butler MS (2008) Natural products to drugs: natural product-derived compounds in clinical trials. *Nat Prod Rep* 25(3):475-516.
62. Zhu F, *et al.* (2011) Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc Natl Acad Sci U S A* 108(31):12943-12948.
63. Vistoli G, Pedretti A, & Testa B (2008) Assessing drug-likeness--what are we missing? *Drug Discov Today* 13(7-8):285-294.
64. Baell JB & Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 53(7):2719-2740.
65. Wang J (2009) Comprehensive assessment of ADMET risks in drug discovery. *Curr Pharm Des* 15(19):2195-2219.
66. Molinski TF, Dalisay DS, Lievens SL, & Saludes JP (2009) Drug development from marine natural products. *Nat Rev Drug Discov* 8(1):69-85.
67. Li JW & Vederas JC (2009) Drug discovery and natural products: end of an era or an endless frontier? *Science* 325(5937):161-165.
68. Gulder TA & Moore BS (2010) Salinosporamide natural products: Potent 20 S proteasome inhibitors as promising cancer chemotherapeutics. *Angew Chem Int Ed Engl* 49(49):9346-9367.
69. Carter GT (2011) Natural products and Pharma 2011: strategic changes spur new opportunities. *Nat Prod Rep* 28(11):1783-1789.
70. Junio HA, *et al.* (2011) Synergy-directed fractionation of botanical medicines: a case study with goldenseal (*Hydrastis canadensis*). *J Nat Prod* 74(7):1621-1629.
71. Gertsch J (2011) Botanical drugs, synergy, and network pharmacology: forth and back to intelligent mixtures. *Planta Med* 77(11):1086-1098.
72. Shabbir M, Love J, & Montgomery B (2008) Phase I trial of PC-Spes2 in advanced hormone refractory prostate cancer. *Oncol Rep* 19(3):831-835.
73. Ma XH, *et al.* (2009) Synergistic therapeutic actions of herbal ingredients and their mechanisms from molecular interaction and network perspectives. *Drug Discov Today* 14(11-12):579-588.
74. Cochrane ZR, Gregory P, & Wilson A (2011) Quality of natural product clinical trials: a comparison of those published in alternative medicine versus conventional medicine journals. *J Diet Suppl* 8(2):135-143.
75. Eisenberg DM, *et al.* (1998) Trends in alternative medicine use in the United States, 1990-1997: results of a follow-up national survey. *JAMA* 280(18):1569-1575.
76. Cordell GA & Colvard MD (2012) Natural products and traditional medicine: turning on a paradigm. *J Nat Prod* 75(3):514-525.
77. Jia J, *et al.* (2009) Mechanisms of drug combinations: interaction and network perspectives. *Nat Rev Drug Discov* 8(2):111-128.
78. Kaiser J (2011) Combining targeted drugs to stop resistant tumors. *Science* 331(6024):1542-1545.
79. Lewith GT, Hyland ME, & Shaw S (2002) Do attitudes toward and beliefs about

- complementary medicine affect treatment outcomes? *American Journal of Public Health* 92(10):1604-1606.
80. Barker Bausell R (2009) Are positive alternative medical therapy trials credible?: Evidence from four high-impact medical journals. *Eval Health Prof* 32(4):349-369.
81. Staud R (2011) Effectiveness of CAM therapy: understanding the evidence. *Rheum Dis Clin North Am* 37(1):9-17.
82. Williamson EM (2001) Synergy and other interactions in phytomedicines. *Phytomedicine* 8(5):401-409.
83. Dinan L, *et al.* (2001) Assessment of natural products in the *Drosophila melanogaster* B(II) cell bioassay for ecdysteroid agonist and antagonist activities. *Cell Mol Life Sci* 58(2):321-342.
84. Stermitz FR, Lorenz P, Tawara JN, Zenewicz LA, & Lewis K (2000) Synergy in a medicinal plant: antimicrobial action of berberine potentiated by 5'-methoxyhydrnocarpin, a multidrug pump inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* 97(4):1433-1437.
85. Wagner H (2011) Synergy research: approaching a new generation of phytopharmaceuticals. *Fitoterapia* 82(1):34-37.
86. Efferth T & Koch E (2011) Complex interactions between phytochemicals. The multi-target therapeutic concept of phytotherapy. *Curr Drug Targets* 12(1):122-132.
87. Knox C, *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 39(Database issue):D1035-1041.
88. Zhu F, *et al.* (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 40(Database issue):D1128-1136.
89. Yang IS, *et al.* (2008) IDBD: infectious disease biomarker database. *Nucleic Acids Res* 36(Database issue):D455-460.
90. Srivastava S (2013) The early detection research network: 10-year outlook. *Clin Chem* 59(1):60-67.
91. Yang W, *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 41(Database issue):D955-961.
92. Bramer GR (1988) International statistical classification of diseases and related health problems. Tenth revision. *World Health Stat Q* 41(1):32-36.
93. Wood PH (1990) Applications of the International Classification of Diseases. *World Health Stat Q* 43(4):263-268.
94. Sayers EW, *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 40(Database issue):D13-25.
95. Averill R & Bowman S (2012) There are critical reasons for not further delaying the implementation of the new ICD-10 coding system. *J AHIMA* 83(7):42-48; quiz 49.
96. Topaz M, Shafran-Topaz L, & Bowles KH (2013) ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. *Perspect Health Inf Manag* 10:1d.
97. Zhang J, *et al.* (2012) Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors. *Mol Biosyst* 8(10):2645-2656.
98. Califano R, Abidin AZ, Peck R, Faivre-Finn C, & Lorigan P (2012) Management of small cell

- lung cancer: recent developments for optimal care. *Drugs* 72(4):471-490.
99. Hilsenbeck SG, *et al.* (1999) Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *Journal of the National Cancer Institute* 91(5):453-459.
100. Staunton JE, *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A* 98(19):10787-10792.
101. Rosenwald A, *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England journal of medicine* 346(25):1937-1947.
102. Takata R, *et al.* (2005) Predicting response to methotrexate, vinblastine, doxorubicin, and cisplatin neoadjuvant chemotherapy for bladder cancers through genome-wide gene expression profiling. *Clinical cancer research : an official journal of the American Association for Cancer Research* 11(7):2625-2636.
103. Balko JM, *et al.* (2006) Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. *BMC genomics* 7:289.
104. Okano T, *et al.* (2007) Proteomic signature corresponding to the response to gefitinib (Iressa, ZD1839), an epidermal growth factor receptor tyrosine kinase inhibitor in lung adenocarcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research* 13(3):799-805.
105. Ma Y, *et al.* (2006) Predicting cancer drug response by proteomic profiling. *Clinical cancer research : an official journal of the American Association for Cancer Research* 12(15):4583-4589.
106. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85-94.
107. Bender A (2010) Databases: Compound bioactivities go public. *Nat Chem Biol* 6(5):309-309.
108. Liu T, Lin Y, Wen X, Jorissen RN, & Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 35(Database issue):D198-201.
109. Yamane S, *et al.* (2008) Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients. *J Inflamm (Lond)* 5:5.
110. Ma XH, *et al.* (2008) Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. *J Chem Inf Model* 48(6):1227-1237.
111. Oprea TI & Gottfries J (2001) Chemography: the art of navigating in chemical space. *J. Comb. Chem* 3(2):157-166.
112. Reymond TFaJ-L (2007) Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* 47(2):342-353.
113. Koch MA, *et al.* (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* 102(48):17272-17277.
114. Han LY, *et al.* (2008) A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate

- and enrichment factor. *J Mol Graph Model* 26(8):1276-1286.
115. Briem H & Gunther J (2005) Classifying "kinase inhibitor-likeness" by using machine-learning methods. *Chembiochem* 6(3):558-566.
116. Glick M, Jenkins JL, Nettles JH, Hitchings H, & Davies JW (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model* 46(1):193-200.
117. Xue Y, *et al.* (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci* 44(4):1497-1505.
118. Tong W, *et al.* (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* 112(12):1249-1254.
119. Ijjaali I, Petitet F, Dubus E, Barberan O, & Michel A (2007) Assessing potency of c-Jun N-terminal kinase 3 (JNK3) inhibitors using 2D molecular descriptors and binary QSAR methodology. *Bioorg Med Chem* 15(12):4256-4264.
120. Vapnik VN (1995) *The nature of statistical learning theory* (Springer, New York).
121. Pochet N, De Smet F, Suykens JA, & De Moor BL (2004) Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* 20:3185-3195.
122. Li F & Yang Y (2005) Analysis of recursive gene selection approaches from microarray data. *Bioinformatics* 21:3741-3747.
123. Johnson RA & Wichern DW (1982) *Applied multivariate statistical analysis* (Prentice Hall, Englewood Cliffs, NJ).
124. Fix E & Hodges JL (1951) *Discriminatory analysis: Non-parametric discrimination: Consistency properties*. (USAF School of Aviation Medicine, Randolph Field, Texas) pp 261-279.
125. Specht DF (1990) Probabilistic neural networks. *Neural Networks* 3(1):109-118.
126. Spitzer R, Cleves AE, & Jain AN (2011) Surface-based protein binding pocket similarity. *Proteins* 79(9):2746-2763.
127. Das S, Kokardekar A, & Breneman CM (2009) Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J Chem Inf Model* 49(12):2863-2872.
128. Hieke M, *et al.* (2011) Discovery and biological evaluation of a novel class of dual microsomal prostaglandin E2 synthase-1/5-lipoxygenase inhibitors based on 2-[(4,6-diphenethoxypyrimidin-2-yl)thio]hexanoic acid. *J Med Chem* 54(13):4490-4507.
129. Tang J, *et al.* (2011) 6-Benzoyl-3-hydroxypyrimidine-2,4-diones as dual inhibitors of HIV reverse transcriptase and integrase. *Bioorg Med Chem Lett* 21(8):2400-2402.
130. Qian L, Liao SY, Huang ZL, Shen Y, & Zheng KC (2010) Theoretical studies on pyrimidine substituent derivatives as dual inhibitors of AP-1 and NF-kappaB. *J Mol Model* 16(6):1139-1150.
131. Henrich S, Feierberg I, Wang T, Blomberg N, & Wade RC (2010) Comparative binding energy analysis for binding affinity and target selectivity prediction. *Proteins* 78(1):135-153.
132. Bajorath J (2008) Computational analysis of ligand relationships within target families. *Curr Opin Chem Biol* 12(3):352-358.
133. Miduturu CV, *et al.* (2011) High-throughput kinase profiling: a more efficient approach

- toward the discovery of new kinase inhibitors. *Chem Biol* 18(7):868-879.
134. Rix U, *et al.* (2007) Chemical proteomic profiles of the BCR-ABL inhibitors imatinib, nilotinib, and dasatinib reveal novel kinase and nonkinase targets. *Blood* 110(12):4055-4063.
 135. Remsing Rix LL, *et al.* (2009) Global target profile of the kinase inhibitor bosutinib in primary chronic myeloid leukemia cells. *Leukemia* 23(3):477-485.
 136. Rix U, *et al.* (2010) A comprehensive target selectivity survey of the BCR-ABL kinase inhibitor INNO-406 by kinase profiling and chemical proteomics in chronic myeloid leukemia cells. *Leukemia* 24(1):44-50.
 137. Goldstein DM, Gray NS, & Zarrinkar PP (2008) High-throughput kinase profiling as a platform for drug discovery. *Nat Rev Drug Discov* 7(5):391-397.
 138. Aronov AM & Murcko MA (2004) Toward a pharmacophore for kinase frequent hitters. *J Med Chem* 47(23):5616-5619.
 139. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32(7):1466-1474.
 140. Letunic I & Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39(Web Server issue):W475-478.
 141. Shultz MD (2013) Setting expectations in molecular optimizations: Strengths and limitations of commonly used composite parameters. *Bioorg Med Chem Lett* 23(21):5980-5991.
 142. Newman DJ & Cragg GM (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 75(3):311-335.
 143. Fantin B, Leggett J, Ebert S, & Craig WA (1991) Correlation between in vitro and in vivo activity of antimicrobial agents against gram-negative bacilli in a murine infection model. *Antimicrob Agents Chemother* 35(7):1413-1422.
 144. Johnson JI, *et al.* (2001) Relationships between drug activity in NCI preclinical in vitro and in vivo models and early clinical trials. *Br J Cancer* 84(10):1424-1431.
 145. Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer* 6(10):813-823.
 146. Chou TC (2006) Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacol Rev* 58(3):621-681.
 147. Kumar N, Afeyan R, Kim HD, & Lauffenburger DA (2008) Multi-Pathway Model Enables Prediction of Kinase Inhibitor Cross-Talk Effects on Migration of Her2-Overexpressing Mammary Epithelial Cells. *Mol Pharmacol*.
 148. Xiong H & Choe Y (2008) Dynamical pathway analysis. *BMC Syst Biol* 2(1):9.
 149. Sivachenko A, Kalinin A, & Yuryev A (2006) Pathway analysis for design of promiscuous drugs and selective drug mixtures. *Curr Drug Discov Technol* 3(4):269-277.
 150. Kim HS & Fay JC (2007) Genetic variation in the cysteine biosynthesis pathway causes sensitivity to pharmacological compounds. *Proc Natl Acad Sci U S A* 104(49):19387-19391.
 151. Carvalho-Netto EF, Markham C, Blanchard DC, Nunes-de-Souza RL, & Blanchard RJ (2006) Physical environment modulates the behavioral responses induced by chemical stimulation of dorsal periaqueductal gray in mice. *Pharmacol Biochem Behav* 85(1):140-147.
 152. Yang H, *et al.* (2007) Nutrient-sensitive mitochondrial NAD⁺ levels dictate cell survival. *Cell* 130(6):1095-1107.
 153. Tabernero J, *et al.* (2008) Dose- and Schedule-Dependent Inhibition of the Mammalian Target of Rapamycin Pathway With Everolimus: A Phase I Tumor Pharmacodynamic Study in

- Patients With Advanced Solid Tumors. *J Clin Oncol*.
154. Oprea TI, *et al.* (2009) A crowdsourcing evaluation of the NIH chemical probes. *Nat Chem Biol* 5(7):441-447.
155. MacIndoe JH, Woods GR, Etre LA, & Covey DF (1982) Comparative studies of aromatase inhibitors in cultured human breast cancer cells. *Cancer research* 42(8 Suppl):3378s-3381s.
156. Quach H, *et al.* (2010) Mechanism of action of immunomodulatory drugs (IMiDs) in multiple myeloma. *Leukemia* 24(1):22-32.
157. Hayashi K, Shimura K, Makino T, & Mizukami H (2010) Comparison of the contents of kampo decoctions containing ephedra herb when prepared simply or by re-boiling according to the traditional theory. *J Nat Med* 64(1):70-74.
158. Zhou XM, *et al.* (2011) In vivo anti-avian influenza virus activity of Qingkailing and Shuanghuanglian Orals. *Chinese Traditional and Herbal Drugs* 42(7):1351-1356.
159. Li LJ, *et al.* (2005) Anti-Virus Activities of the Extract and Effective Components Isolated from *Senecio Cannabifolius* Less. *Chinese Journal of Basic Medicine in Traditional Chinese Medicine* 11(8):585-587.
160. Csermely P, Agoston V, & Pongor S (2005) The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol Sci* 26(4):178-182.
161. Xie L, Evangelidis T, & Bourne PE (2011) Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput Biol* 7(4):e1002037.
162. Wang L, *et al.* (2008) Dissection of mechanisms of Chinese medicinal formula Realgar-Indigo naturalis as an effective treatment for promyelocytic leukemia. *Proc Natl Acad Sci U S A* 105(12):4826-4831.
163. Clark KJ, *et al.* (1998) An in vitro study of theaflavins extracted from black tea to neutralize bovine rotavirus and bovine coronavirus infections. *Veterinary microbiology* 63(2-4):147-157.
164. Papp B, Pal C, & Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429(6992):661-664.
165. Tong X, *et al.* (2004) TCM treatment of infectious atypical pneumonia--a report of 16 cases. *J Tradit Chin Med* 24(4):266-269.
166. Zhang GG, *et al.* (2005) Variability in the traditional Chinese medicine (TCM) diagnoses and herbal prescriptions provided by three TCM practitioners for 40 patients with rheumatoid arthritis. *J Altern Complement Med* 11(3):415-421.
167. Larder BA, Kemp SD, & Harrigan PR (1995) Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy. *Science* 269(5224):696-699.
168. Dancey JE & Chen HX (2006) Strategies for optimizing combinations of molecularly targeted anticancer agents. *Nat Rev Drug Discov* 5(8):649-659.
169. Silver LL (2007) Multi-targeting by monotherapeutic antibacterials. *Nat Rev Drug Discov* 6(1):41-55.
170. Chiu PH, Hsieh HY, & Wang SC (2012) Prescriptions of traditional Chinese medicine are specific to cancer types and adjustable to temperature changes. *PLoS One* 7(2):e31648.
171. Efferth T (2010) Personalized cancer medicine: from molecular diagnostics to targeted therapy with natural products. *Planta Med* 76(11):1143-1154.
172. Harvey AL & Cree IA (2010) High-throughput screening of natural products for cancer

- therapy. *Planta Med* 76(11):1080-1086.
173. Georgakis GV, Li Y, Rassidakis GZ, Medeiros LJ, & Younes A (2006) The HSP90 inhibitor 17-AAG synergizes with doxorubicin and U0126 in anaplastic large cell lymphoma irrespective of ALK expression. *Exp Hematol* 34(12):1670-1679.
174. Kassouf W, *et al.* (2005) Uncoupling between epidermal growth factor receptor and downstream signals defines resistance to the antiproliferative effect of Gefitinib in bladder cancer cells. *Cancer research* 65(22):10524-10535.
175. Thanou M, Verhoef JC, & Junginger HE (2001) Oral drug absorption enhancement by chitosan and its derivatives. *Adv Drug Deliv Rev* 52(2):117-126.
176. Robinson WH, Lindstrom TM, Cheung RK, & Sokolove J (2013) Mechanistic biomarkers for clinical decision making in rheumatic diseases. *Nat Rev Rheumatol* 9(5):267-276.
177. Ludwig JA & Weinstein JN (2005) Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer* 5(11):845-856.
178. Walsh P, Elsabbagh M, Bolton P, & Singh I (2011) In search of biomarkers for autism: scientific, social and ethical challenges. *Nat Rev Neurosci* 12(10):603-612.
179. Ho C & Laskin J (2009) EGFR-directed therapies to treat non-small-cell lung cancer. *Expert Opin Investig Drugs* 18(8):1133-1145.
180. Linardou H, Dahabreh IJ, Bafaloukos D, Kosmidis P, & Murray S (2009) Somatic EGFR mutations and efficacy of tyrosine kinase inhibitors in NSCLC. *Nat Rev Clin Oncol* 6(6):352-366.
181. Blennow K, Hampel H, & Zetterberg H (2013) Biomarkers in Amyloid-beta Immunotherapy Trials in Alzheimer's Disease. *Neuropsychopharmacology*.
182. Cancer Genome Atlas N (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418):61-70.
183. Hauschild A, *et al.* (2012) Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. *Lancet* 380(9839):358-365.
184. Catalanotti F & Solit DB (2012) Will Hsp90 inhibitors prove effective in BRAF-mutant melanomas? *Clinical cancer research : an official journal of the American Association for Cancer Research* 18(9):2420-2422.
185. Cichowski K & Janne PA (2010) Drug discovery: inhibitors that activate. *Nature* 464(7287):358-359.
186. Nazarian R, *et al.* (2010) Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* 468(7326):973-977.
187. Li QL, *et al.* (2012) Activation of PI3K/AKT and MAPK pathway through a PDGFRbeta-dependent feedback loop is involved in rapamycin resistance in hepatocellular carcinoma. *PLoS One* 7(3):e33379.
188. Cools J, *et al.* (2004) Prediction of resistance to small molecule FLT3 inhibitors: implications for molecularly targeted therapy of acute leukemia. *Cancer research* 64(18):6385-6389.
189. Rizvi NA, *et al.* (2011) Molecular characteristics predict clinical outcomes: prospective trial correlating response to the EGFR tyrosine kinase inhibitor gefitinib with the presence of sensitizing mutations in the tyrosine binding domain of the EGFR gene. *Clinical cancer research : an official journal of the American Association for Cancer Research* 17(10):3500-3506.
190. Molinari F, *et al.* (2011) Increased detection sensitivity for KRAS mutations enhances the

- prediction of anti-EGFR monoclonal antibody resistance in metastatic colorectal cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* 17(14):4901-4914.
191. Wang W, Cassidy J, O'Brien V, Ryan KM, & Collie-Duguid E (2004) Mechanistic and predictive profiling of 5-Fluorouracil resistance in human cancer cells. *Cancer research* 64(22):8167-8176.
192. Dai Z, Barbacioru C, Huang Y, & Sadee W (2006) Prediction of anticancer drug potency from expression of genes involved in growth factor signaling. *Pharm Res* 23(2):336-349.
193. Brase JC, et al. (2010) ERBB2 and TOP2A in breast cancer: a comprehensive analysis of gene amplification, RNA levels, and protein expression and their influence on prognosis and prediction. *Clinical cancer research : an official journal of the American Association for Cancer Research* 16(8):2391-2401.
194. Steinhubl SR, Muse ED, & Topol EJ (2013) Can mobile health technologies transform health care? *JAMA* 310(22):2395-2396.
195. Sieverdes JC, Treiber F, & Jenkins C (2013) Improving diabetes management with mobile health technology. *The American journal of the medical sciences* 345(4):289-295.
196. Kouris I, Tsirmpas C, Mouggiakakou SG, Iliopoulou D, & Koutsouris D (2010) E-Health towards ecumenical framework for personalized medicine via Decision Support System. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 2010:2881-2885.
197. Bruderman I & Abboud S (1997) Telespirometry: novel system for home monitoring of asthmatic patients. *Telemedicine journal : the official journal of the American Telemedicine Association* 3(2):127-133.
198. Liao O, Morpew T, Amaro S, & Galant SP (2006) The Breathmobile: a novel comprehensive school-based mobile asthma care clinic for urban underprivileged children. *The Journal of school health* 76(6):313-319.
199. Lee RG, Chen KC, Hsiao CC, & Tseng CL (2007) A mobile care system with alert mechanism. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society* 11(5):507-517.
200. Lee YG, Jeong WS, & Yoon G (2012) Smartphone-based mobile health monitoring. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* 18(8):585-590.
201. Piette JD, et al. (2012) Hypertension management using mobile technology and home blood pressure monitoring: results of a randomized trial in two low/middle-income countries. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* 18(8):613-620.
202. Stuckey MI, Shapiro S, Gill DP, & Petrella RJ (2013) A lifestyle intervention supported by mobile health technologies to improve the cardiometabolic risk profile of individuals at risk for cardiovascular disease and type 2 diabetes: study rationale and protocol. *BMC public health* 13:1051.
203. Perez F, et al. (2006) Evaluation of a mobile health system for supporting postoperative patients following day surgery. *Journal of telemedicine and telecare* 12 Suppl 1:41-43.
204. Majewski IJ & Bernards R (2011) Taming the dragon: genomic biomarkers to individualize

- the treatment of cancer. *Nature medicine* 17(3):304-312.
205. Maisel AS & Choudhary R (2012) Biomarkers in acute heart failure--state of the art. *Nat Rev Cardiol* 9(8):478-490.
 206. Blennow K (2010) Biomarkers in Alzheimer's disease drug development. *Nature medicine* 16(11):1218-1222.
 207. Sreekumar A, *et al.* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 457(7231):910-914.
 208. Wang S, *et al.* (2011) Integration of cell phone imaging with microchip ELISA to detect ovarian cancer HE4 biomarker in urine at the point-of-care. *Lab on a chip* 11(20):3411-3418.
 209. Mendes B, Silva P, Aveiro F, Pereira J, & Camara JS (2013) A micro-extraction technique using a new digitally controlled syringe combined with UHPLC for assessment of urinary biomarkers of oxidatively damaged DNA. *PLoS One* 8(3):e58366.
 210. Hsu MY, *et al.* (2014) Monitoring the VEGF level in aqueous humor of patients with ophthalmologically relevant diseases via ultrahigh sensitive paper-based ELISA. *Biomaterials* 35(12):3729-3735.
 211. Costa MN, *et al.* (2014) A low cost, safe, disposable, rapid and self-sustainable paper-based platform for diagnostic testing: lab-on-paper. *Nanotechnology* 25(9):094006.
 212. Yamada K, Takaki S, Komuro N, Suzuki K, & Citterio D (2014) An antibody-free microfluidic paper-based analytical device for the determination of tear fluid lactoferrin by fluorescence sensitization of Tb³⁺. *The Analyst* 139(7):1637-1643.
 213. Fobel R, Kirby AE, Ng AH, Farnood RR, & Wheeler AR (2014) Paper microfluidics goes digital. *Advanced materials* 26(18):2838-2843.
 214. Vashist SK, Mudanyali O, Schneider EM, Zengerle R, & Ozcan A (2014) Cellphone-based devices for bioanalytical sciences. *Analytical and bioanalytical chemistry* 406(14):3263-3277.
 215. Warren AD, Kwong GA, Wood DK, Lin KY, & Bhatia SN (2014) Point-of-care diagnostics for noncommunicable diseases using synthetic urinary biomarkers and paper microfluidics. *Proc Natl Acad Sci U S A* 111(10):3671-3676.
 216. Bourquin Y, Reboud J, Wilson R, Zhang Y, & Cooper JM (2011) Integrated immunoassay using tuneable surface acoustic waves and lensfree detection. *Lab on a chip* 11(16):2725-2730.
 217. Cheng CM, *et al.* (2010) Paper-based ELISA. *Angew Chem Int Ed Engl* 49(28):4771-4774.
 218. Stedtfeld RD, *et al.* (2012) Gene-Z: a device for point of care genetic testing using a smartphone. *Lab on a chip* 12(8):1454-1462.
 219. Brower V (2011) Biomarkers: Portents of malignancy. *Nature* 471(7339):S19-21.
 220. Cote RA & Robboy S (1980) Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA* 243(8):756-762.
 221. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32(Database issue):D267-270.
 222. Qin C, *et al.* (2014) Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res* 42(Database issue):D1118-1123.

Appendices

Supplementary Table S4.1: Targets and potency-enhancing synergistic molecular modes in 3 fully or partially sub-potent natural product combinations with group potencies improved to drug levels.

Ingredient	Role in Combination	Dose Reduction Index	Therapeutic Effect or Response	Effect type	Synergistic Action for Effect or Against Negative Response	Type of Synergism
Tetraarsenic tetrasulfide (1.1uM)	Main therapeutic component	6.88	degraded the PML-RAR oncoprotein leading to anticancer effect (18344322), which may also hinder the possible antagonistic effect of RAR on TGF-alpha induced growth inhibition (12527889)	Anticancer, growth inhibition, via RAR	Indirubin blocked VEGFR2-mediated JAK/STAT3 signaling (21207415), which partially hindered RAR-STAT3 crosstalk and its promotion of apoptosis resistance (14959844) and transcription activation (15044588), thereby enhancing tetraarsenic tetrasulfide's anticancer effect	Complementary action
					Tanshinone IIA reduced and antagonized androgen receptor (AR) (22175694, 22281759, 21997969) to hinders its effect on the upregulation of RAR (12069693), thereby adding on	Complementary action

Appendices

					tetraarsenic tetrasulfide's action in reducing PML-RAR	
			down-regulated CDK2 in NB4 and NB4-R2 cells (18344322)	Anticancer, cell cycle regulation	Indirubin inhibited and down-regulated CDK2 (18344322) to complement tetraarsenic tetrasulfide's action in reducing CDK2	Complementary action
			upregulated RING-type E3 ligase c-CBL, leading to degradation of BCR-ABL (21118980)	Anti-leukemia, growth inhibition		
			tetraarsenic tetrasulfide transported into tumor cell by AQP9 (18344322)	bioavailability	Indirubin upregulated APQ9 (18344322) to enhance Tetraarsenic tetrasulfide's bioavailability	bioavailability enhancement
					Tanshinone IIA (T) upregulated APQ9 (18344322) to enhance Tetraarsenic tetrasulfide's bioavailability	bioavailability enhancement
			Reduction in RAR alpha may lead to P53 downregulation and Bcl-2 upregulation, thereby countering anticancer effect (10675490)	Counteractive action against anticancer effect	Tanshinone IIA's activation of p53 signaling (21997969) may help against this counteractive action	Anti-counteractive action
Indirubin (>3uM)	Cooperative	>9.38	inhibited and down-regulated CDK2, leading to anticancer effect (18344322)	Anticancer, cell cycle regulation	Tetraarsenic tetrasulfide reduced CDK2 (18344322), thereby complementing	Complementary action

					indirubin's action on CDK2	
			inhibited GSK-3, thereby blocked its effect on tumor proliferation and migration (21697283)	Anticancer, growth inhibition		
			inhibited angiogenesis via blocking VEGFR2-mediated JAK/STAT3 signaling (21207415), which also partially hindered RAR-STAT3 crosstalk and its action on apoptosis resistance (14959844) and transcription activation (15044588), leading to anticancer effect	Anticancer, growth and angiogenesis inhibition, via RAR partner and additional growth and angiogenesis signaling		
			activated AhR (20951181), and AhR activation may lead to the activation of RAR alpha in the absence of ligand, thereby countering anticancer effect (16480812)	Counteractive action against anticancer effect, via RAR regulator	Tetraarsenic tetrasulfide degraded the PML-RAR oncoprotein (18344322), which may patially alleviate the conteractive action	Anti-counteractive action
Tanshinone IIA (>3uM)	Cooperative	>9.38	increased Bax/Bcl-2 ratio and caspase 3, decreased Bcl-2, mitochondrial membrane potential, MMPs and CD31, leading to anticancer apoptosis effects (21472292, 22002472, 22126901)	Anticancer, apoptosis		

			activated p53 signaling to promote anticancer effect (21997969)	Anticancer, cell cycle regulation, apoptosis		
			reduced survivin, ERCC1 and LRP via upregulation of phospho-P38, leading to enhanced apoptosis and anticancer effect (21165580)	Anticancer, apoptosis		
			reduced HER2, NF-κBp65 and LC3-II, leading to anticancer effect particularly against breast cancer (22246196), NF-κBp65 downregulation further hindered the effect of the binding of NF-κBp65 and RAR alpha on transcriptional regulation (17451432), which further contributes the ingredient's anticancer effects	Anticancer, apoptosis, growth inhibition, via RAR regulator and additional growth and survival signaling		
			reduced and antagonized androgen receptor (AR) and induced apoptosis, leading to anticancer effect against prostate cancer (22175694, 22281759, 21997969)	Anticancer, growth inhibition		
			upregulated phospho-P38 (21165580), which may help directing RAR alpha to its target promoters (19078967) and to	Counteractive action against anticancer effect		

			subsequently cooperate with other cancer proteins for effective transcriptional activity in certain cancers (20080953), thereby countering anticancer effect			
			increased efflux transporters, which may help effluxing the ingredient (a Pgp substrate), thereby lowering its bioavailability (17504222, 20821829)	Efflux-mediated multidrug resistance	Indirubin may inhibit certain efflux pump (20380543) to reduce the efflux of Tanshinone IIA	bioavailability enhancement
Theaflavin (0.943ug/mL)	Main therapeutic component	9.33	Rotavirus activated JNK and p38 signaling pathways for enhanced viral replication (16928761), theaflavin reduced JNK and P38 phosphorelation (21184129, 22111069), which hinders viral replication and leads to virus neutralisation	Antiviral, against two of the 4 redundant viral replication regulators	The four ingredients target 4 redundant viral replication regulators, leading to strong synergistic antiviral activity, such activity is further enhanced by pathways that mediate viral survival, growth and cell entry	Complementary action against redundant regulators
theaflavin-3-monogallate (251.39ug/mL)	Cooperative	2489	Rotavirus activated Cox2 to mediate viral infection at a postbinding step (15331705) probably including viral replication (17555580), theaflavin-3-monogallate and theaflavin-3'-monogallate mixture downregulated Cox2 (11103814),	Antiviral, against one of the 4 redundant viral replication regulators		

Appendices

			which hinders viral replication and leads to virus neutralisation			
theaflavin-3'-monogallate (5.07ug/mL)	Cooperative	50.2	Rotavirus activated Cox2 to mediate viral infection at a postbinding step (15331705) probably including viral replication (17555580), theaflavin-3-monogallate and theaflavin-3'-monogallate mixture downregulated Cox2 (11103814), which hinders viral replication and leads to virus neutralisation	Antiviral, against one of the 4 redundant viral replication regulators		
theaflavin-3,3' digallate (5.51ug/mL)	Cooperative	54.6	Rotavirus activated ERK signaling pathways for enhanced viral replication (17689685), theaflavin-3,3' digallate reduced ERK phosphorelation (11511526), which hinders viral replication and leads to virus neutralisation	Antiviral, against one of the 4 redundant viral replication regulators		
			Rotavirus activated NFkB and AKT signaling pathways to suppress virus-induced cellular apoptosis and facilitate viral growth (20392855), theaflavin-3,3'-digallate blocked	Antiviral, against viral survival and growth		

			NFkB activation (16880762) thereby hindered viral growth			
All four ingredients			Rotavirus's entry into cells is partly facilitated by integrin, and rotavirus replication upregulated alpha2beta1 and beta2 integrins via activation of PI3K pathways, leading to further enhanced viral entry (17942548), theaflavins educed PI3K and pAkt (14743383), thereby reduced the enhancement of viral entry	Antiviral, against viral entry		
			Possible Antiviral Effect			
theaflavin-3,3' digallate (5.51ug/mL)			Rotavirus's entry into cells is partly facilitated by haemagglutinin (15165605), theaflavin-3,3' digallate inhibited haemagglutinin of influenza virus (8215301), if it also inhibits haemagglutinin of rotavirus, theaflavin-3,3' digallate may hinder the viral entry process leading to virak neutralisation	Antiviral, against viral entry		

wedelolactone (0.8uM)	Main therapeutic component	63.5	Potent androgen receptor (AR) antagonist (IC50 0.2uM) (17942463)	Anticancer, growth inhibitin, via AR	indole-3-carboxylaldehyde's assumed AR downregulation (17942463) complements wedelolactone's AR antagonism, leading to synergism	Complementary action
					luteolin reduced AR expression in dose and time dependent manner (18008333), which complements wedelolactone's AR antagonism, leading to synergism	Complementary action
					luteolin inhibited c-Src activities (20215519), which hinders c-Src mediated enhancement of AR activity and AR transactivation function (21135112, 18223692), thereby complementing wedelolactone's AR antagonism, leading to synergism	Complementary action
					luteolin downregulated FGF1R signaling (22269172) to hinder its crosstalk with AR and stimulation of AR activity (21465482), which complements wedelolactone's AR antagonism, leading to synergism	Complementary action

					luteolin inhibited topoisomerase II (19149659) to hinder AR and topoisomerase II beta binding mediated oncogenic rearrangement and DNA repair (20601956. 21385925), thereby complementing wedelolactone's AR antagonism, leading to synergism	Complementary action
					apigenin inhibited CK2 (21871133), CK2 inhibition lead to AR downregulation in prostate cancer cells (17044081), thereby complementing wedelolactone's AR antagonism, leading to synergism	Complementary action
					apigenin reduced EGFR and HER2 expression (21196218) to hinder EGFR and HER2 mediated AR activation and prostate cancer progression (19318561), thereby complementing wedelolactone's AR antagonism, leading to synergism	Complementary action
					apigenin inhibited NF- κ B activation (21142820) to hinder its promotion of AR expression in prostate cancer cells (18701501, 19628766), thereby complementing wedelolactone's AR	Complementary action

Appendices

					antagonism, leading to synergism	
					apigenin activated P53 (22227579), which helps downregulating AR in prostate cancer (18084622)	Complementary action
					luteolin and apigenin each suppressed Akt activation (20655656, 22084167), which hinders AkT-mediated AR upregulation in prostate cancer (21317204), thereby complementing wedelolactone's AR antagonism, leading to synergism	Complementary action
					luteolin and apigenin each reduced CDK6 activity (20655656, 16648554), which hinders CDK6-mediated enhancement of AR transcriptional activity in prostate cancer cells (15790678) to complement AR antagonism, leading to synergism	Complementary action

Appendices

					luteolin and apigenin each inhibited GSK-3 β (IC50 1.5 and 1.9 μ M) (21443429), GSK-3 β inhibition helps AR export from cell nucleus thereby diminishing its effects (21980429), which complements AR antagonism, leading to synergism	Complementary action
					luteolin and apigenin each inhibited HDAC (IC50 50-100 and 20-40 μ M) (21074525, 22006862) to hinder HDAC's facilitating action on AR function in hormone-sensitive and castrate-resistant prostate cancer (19176386), thereby complementing wedelolactone's AR antagonism, leading to synergism	Complementary action
			inhibited the activity of DNA topoisomerase α independent of AR (21315506)	Anticancer, apoptosis	.	
			inhibited IKK leading to reduced activation of Akt and NF κ B (21704149)	Anticancer, growth inhibition and apoptosis	.	
			inhibited trypsin (12722155), trypsin may have tumor suppressive activity (14583448) and this activity may be	Counteractive action against anticancer effect	apigenin activated P53 (22227579), which helps downregulating AR in prostate cancer (18084622)	Complementary action

			hindered by wedelolactone			
indole-3-carboxylaldehyde (656uM)	Cooperative	15238	indole-3-carboxylaldehyde's structural analog indole-3-carbinol (I3C) forms DIM under acidic conditions; with a more stable and more potent anticancer activity, I3C and DIM antagonized androgen binding to AR and down-regulated AR in PCa cells (17942463). It has been speculated that indole-3-carboxylaldehyde may have similar activities (17942463)	Anticancer, growth inhibition, likely via AR down-regulation and antagonism	.	
luteolin (1.72uM)	Cooperative	3.13	AR antagonist (IC50 2.4uM) (17942463)	Anticancer, growth inhibition, via AR antagonism	indole-3-carboxylaldehyde's assumed AR downregulation (17942463) complements luteolin's AR antagonism, leading to synergism	Complementary action
			reduced AR expression in dose and time dependent manner (18008333), which complements its AR antagonist activity	Anticancer, growth inhibition, via AR reduction	.	

			suppressed Akt phosphorylation and activation (20655656), AR levels are upregulated by Akt in prostate cancer (21317204) and AR induces prostate cancer cell proliferation through mTOR activation (16885382), luteolin thus hinders Akt-mediated proliferation and survival, and helps containing AR levels	Anticancer, growth inhibition and apoptosis, via AR regulator and additional proliferation and survival signaling	.	
			reduced CDK4/6 activity (20655656), CDK6 associates with AR and enhances its transcriptional activity in prostate cancer cells (15790678), luteolin may thus hinders CDK4/6 mediated cell-cycle progression and AR expression	Anticancer, growth inhibition and cell cycle regulation, via dual AR and cell cycle regulator	.	
			inhibited c-Src activities (20215519), which hinders c-Src mediated growth signaling (18045060) and enhancement of AR activity and AR transactivation function (21135112, 18223692)	Anticancer, growth inhibition, via AR regulator and additional proliferation signaling	.	
			GSK-3 β inhibitor (IC50 1.5uM) (21443429), GSK-3 β inhibition helps AR export from cell nucleus thereby	Anticancer, growth inhibition, via AR regulator	.	

			diminishing its effects (21980429), which complements AR antagonism			
			HDAC inhibitor (IC50 50-100uM) (21074525), which hinders HDAC's transcription regulation (19383284) and facilitating action on AR function in hormone-sensitive and castrate-resistant prostate cancer (19176386), thereby complementing AR antagonism	Anticancer, growth inhibition, via AR regulator and additional growth regulation	.	
			downregulated FGF1R signaling (22269172), which hinders its growth and survival signaling (19183230) and its crosstalk with AR and stimulation of AR activity (21465482)	Anticancer, growth inhibition, via AR regulator and additional growth and survival regulation	.	
			inhibited topoisomerases I and II (19149659), which hinders AR and topoisomerase II beta binding mediated oncogenic rearrangement and DNA repair (20601956. 21385925)	Anticancer, growth and survival inhibition, via AR partner and additional growth and survival regulation	.	
			activated AMPK (21468539), AMPK	Anticancer, growth	.	

Appendices

			activation reduced growth of prostate cancer cells (19347029)	inhibition		
			inhibited proteasome (22292765), which may induce apoptosis (18347166)	Anticancer, apoptosis	.	
			downregulated Cyclin D1 (20655656), AR is strongly suppressed by cyclin D1 (21212260), luteolin may thus both hinder cell cycle and reduce AR suppression	Anticancer cell cycle regulation	.	
				Counteractive action against anticancer effect	.	
			activated P38 (22073986, 21762691), P38 activation may facilitate androgen-independent AR activation (19151763), which counters luteolin's AR antagonistic effect	Counteractive action against anticancer effect	.	
apigenin (3.02uM)	Cooperative	250	AR antagonist (IC50 9.8uM) (17942463)	Anticancer, growth inhibition, via AR antagonism	indole-3-carboxylaldehyde's assumed AR downregulation (17942463) complements apigenin's AR antagonism, leading to synergism	Complementary action

					luteolin reduced AR expression in dose and time dependent manner (18008333), which complements apigenin's AR antagonism, leading to synergism	Complementary action
			inhibited CK2 (21871133), CK2 inhibition lead to AR downregulation in prostate cancer cells (17044081), thereby complementing its AR antagonism activity	Anticancer, growth inhibition, via AR reduction		
			reduced EGFR and HER2 expression (21196218) to hinder EGFR and HER2 mediated AR activation and prostate cancer progression (19318561) and HER2 signaling in prostate cancer (15769631)	Anticancer, growth inhibition, via AR regulator and additional growth signaling inhibition		
			inactivated Akt (22084167), AR levels are upregulated by Akt in prostate cancer (21317204) and AR induces prostate cancer cell proliferation through mTOR activation (16885382), luteolin thus hinders AkT-mediated proliferation and survival, and helps containing AR levels	Anticancer, growth inhibition and apoptosis, via AR regulator and additional proliferation and survival signaling		

			reduced CDK2/4/6 activity (16648554), CDK6 associates with AR and enhances its transcriptional activity in prostate cancer cells (15790678), apigenin may thus hinders CDK4/6 mediated cell-cycle progression and AR expression	Anticancer, growth inhibition and cell cycle regulation, via dual AR and cell cycle regulator		
			GSK-3 β inhibitor (IC50 1.9uM) (21443429), GSK-3 β inhibition helps AR export from cell nucleus thereby diminishing its effects (21980429), which complements AR antagonism	Anticancer, growth inhibition, via AR regulator		
			HDAC inhibitor (IC50 20-40uM) (22006862), which hinders HDAC's transcription regulation (19383284) and facilitating action on AR function in hormone-sensitive and castrate-resistant prostate cancer (19176386), thereby complementing AR antagonism	Anticancer, growth inhibition, via AR regulator and additional growth regulation		
			inhibited NF- κ B activation (21142820) to hinder its promotion of AR expression in prostate cancer cells (18701501, 19628766), thereby	Anticancer, growth inhibition, via AR regulator		

			complimenting its AR antagonistic effect			
			activated P53 (22227579), which enhances P53 mediated tumor suppressive activity in prostate cancer (21227058[]), and helps downregulating AR in prostate cancer (18084622) and compensating for the reduced P53 activation due to AR downregulation	Anticancer, growth inhibition, via AR regulator and additional tumor suppression activity		
			downregulated Cox-2 expression (20691240) to hinder Cox-2's promotion of prostate cancer progression (12386924)	Anticancer, growth inhibition		
			increased Bax/Bcl-2 ratio, caspase 3 and cytochrome C, decreased Bcl-2, leading to anticancer apoptosis effects (20937639)	Anticancer, apoptosis		
			activated AMPK (21538580), AMPK activation reduced growth of prostate cancer cells (19347029)	Anticancer, growth inhibition		
			upregulated leptin receptor to induce apoptosis (21550230)	Anticancer, apoptosis		
			inhibited proteasome (22292765), which may induce apoptosis			

			(18347166)			
			induced Hsp27 phosphorelation (21364669), Hsp27 mediated repression of AR function in prostate cancer cells (19767773), but promoted IGF1R survival signaling in prostate cancer (20197463), which both complements and counters AR antagonistic activity	Anticancer, growth inhibition, via AR regulator		
				Counteractive action against anticancer effect		
			enhanced P38 phosphorelation (21615506), P38 activation may facilitate androgen-independent AR activation (19151763), which counters apigenin's AR antagonistic effect	Counteractive action against anticancer effect		

Supplementary Table 4.2: Targets and potency-enhancing molecular interaction modes in 2 fully sub-potent natural product combinations with potencies of the principal component increased by >100 fold.

Ingredient	Role in Combination	Dose Reduction Index	Theapeutic Effect or Response	Effect type	Synergistic Action for Effect or Against Negative Response	Synergy Type
aescin (316ug/mL)	main therapeutic component	158	disrupted membrane after metabolism by glycosidases (1171670), leading to haemolysis (21968386)	haemolysis	thymol affected cell membrane structure and enhanced permeability by generating asymmetries and membrane tensions (21660740), thereby facilitating the membrane insertion or crossing of aescin and its subsequent metabolism by glycosidases located in the internal side of membrane (15340929)	bioavailability enhancement
thymol	sensitizer		affected cell membrane structure and enhanced permeability by generating asymmetries and membrane tensions (21660740)	permeability enhancement	.	
<i>n</i>-butylidenephthalide (44.59ug/mL)	main therapeutic component	343	induced orphan nuclear receptor Nur77 to promote apoptosis (18577687, 21365711)	anticancer, apoptosis	.	
			suppressed human telomerase reverse transcriptase to restrict tumor growth (21553143)	anticancer, growth control	.	

			inhibited angiogenesis partly by activating p38 and ERK (21327473)	anticancer, anti-angiogenesis	.	
			induced Nur77 reexpression (18577687, 21365711) may lead to enhanced NFkB activity to reduce apoptosis (16082387), and to induce human telomerase reverse transcriptase for promoting tumor growth (15226182)	counteractive action against anticancer effect	z-ligustilide inhibited NFkB (20581853) to counter this counteractive action	anti-counteractive action
			promoted PI3K-Nurf2 crosstalk to enhance tumor survival signaling	counteractive action against anticancer effect	.	
			reduced P53 to reduce apoptosis (21398513)	counteractive action against anticancer effect	.	
senkyunolide A (10.4ug/mL)	Cooperative	347	anticancer mechanism unreported			
z-ligustilide (11.52ug/mL)	Cooperative	1.92	anticancer mechanism unreported			

Supplementary Table S4.3: Targets and potency-enhancing molecular interaction modes in 9 fully sub-potent natural product combinations with potencies of the principal component increased by 10–100 fold.

Ingredient	Role in Combination	Dose Reduction Index	Theapeutic Effect or Response	Effect type	Synergistic Action for Effect or Against Negative Response	Type of Synergism
------------	---------------------	----------------------	-------------------------------	-------------	--	-------------------

Quillaja saponins (157ug/mL)	main therapeutic component	14.3	induced haemolysis by inserting into and forming pores in membrane, and altering Ca-K and Ca-Mn ATPase activities, and these activities are enhanced by its metabolism by glycosidases (19915999)	haemolysis	affected cell membrane structure and enhanced permeability of anticancer agents by generating asymmetries and membrane tensions (21660740), thereby facilitating membrane insertion or crossing of Quillaja saponins and their subsequent metabolism by glycosidases located in the internal side of membrane (15340929)	bioavailability enhancement
thymol	sensitizer		affected cell membrane structure and enhanced permeability by generating asymmetries and membrane tensions (21660740)	permeability enhancement	.	
Salicylaldehyde (141ug/mL)	main therapeutic component	>26	inhibited fungal antioxidant system proteins cytosolic superoxide dismutase, mitochondrial SOD and glutathione reductase, thereby producing antifungal	antifungal	Linalool increased ROS species in certain cells by (19428344), which complements Salicylaldehyde's inhibition of fungal antioxidant system	complementary action

			effect (20803256)			
Linalool (281ug/mL)	Cooperative	>78	inhibited mitochondrial complexes and increased ROS species in certain cells by (19428344), both inhibition of mitochondrial function and ROS production may contribute to antifungal activity (16834605, 20803256)	antifungal		
berberine (256ug/mL)	main therapeutic component	16	inhibited microbial division protein FtsZ to produce antimicrobial effect (18275156, 21060782)	antimicrobial		
			effluxed by a multidrug pump (10677479)	Efflux-mediated multidrug resistance	5'-methoxyhydnocarpin inhibited the multidrug pump, thereby potentiated berberine's antimicrobial activity (10677479)	bioavailability enhancement

Appendices

5'-methoxyhydnocarpin (10ug/mL)	sensitizer			potentiation		
Linoleic acid (1mg/mL)	co-therapeutic ingredient	20	inhibited bacterial enoyl-acyl carrier protein reductase FabI involved in fatty acid synthesis, thereby producing antibacterial effect (16146629, 21862391)	Antibacterial		
			effluxed by a farAB-encoded bacterial efflux pump (10447892)	Efflux-mediated multidrug resistance	Combination of Linoleic acid and Oleic acid inhibited bacterial efflux (21194895)	bioavailability enhancement
			resisted by bacterial cell wall-anchored proteins SasF and SssF	Counteractive action		
Oleic acid (1mg/mL)	co-therapeutic ingredient	20	inhibited bacterial enoyl-acyl carrier protein reductase FabI involved in fatty acid synthesis, thereby producing antibacterial effect (16146629, 21862391)	Antibacterial		
			effluxed by a farAB-encoded bacterial efflux pump (10447892)	Efflux-mediated multidrug resistance	Combination of Linoleic acid and Oleic acid inhibited bacterial efflux (21194895)	bioavailability enhancement

eriodictyol (0.8mg/mL)	main therapeutic component	16.7	produced prooxidative DNA damage effect (19941260), which may contribute to antimicrobial activity	Antimicrobial	hesperetin has higher bioavailability (20447374) and is converted into eriodictyol in microbial culture by microbial enzymes (21873058), thereby enhancing the bioavailability of eriodictyol	bioavailability enhancement
hesperetin (1mg/mL)	Cooperative	3.33	reduced reducing the activity of bacterial enzymes (18812032), which may contribute to antimicrobial activity	Antimicrobial		
			interacted with membrane better, leading to higher bioavailability inside cells (20447374)	Bioavailability		
eriodictyol (0.8mg/mL)	main therapeutic component	16.7	produced prooxidative DNA damage effect (19941260), which may contribute to antimicrobial activity	Antimicrobial	Naringenin has higher bioavailability (2753859, 7603409, 8132524) and is converted into eriodictyol in microbial culture by microbial enzymes (21299115),	bioavailability enhancement

					thereby enhancing the bioavailability of eriodictyol	
Naringenin (1mg/mL)	Cooperative	2.5	inhibited VacA vacuolation (15770537) to hinder the release of nutrients necessary for microbial growth and survival (12814772), leading to antimicrobial activity	Antimicrobial		
			interacted with membrane better (2753859), leading to higher bioavailability inside cells (7603409, 8132524)	Bioavailability		
Berberine (500ug/mL)	main therapeutic component	16	inhibited microbial division protein FtsZ to produce antimicrobial effect (18275156, 21060782)	antimicrobial		
			effluxed by a multidrug pump (10677479)	Efflux-mediated multidrug resistance	biochanin A inhibited the multidrug pump, thereby potentiated berberine's antimicrobial activity	bioavailability enhancement

					(12952418)	
biochanin A (>312.5ug/mL)	Cooperative	>31.3	inhibited microbial growth (20335979, 21328137)	antimicrobial		
Berberine (500ug/mL)	main therapeutic component	16	inhibited microbial division protein FtsZ to produce antimicrobial effect (18275156, 21060782)	antimicrobial		
			effluxed by a multidrug pump (10677479)	Efflux-mediated multidrug resistance	Genistein inhibited the multidrug pump, thereby potentiated berberine's antimicrobial activity (12952418)	bioavailability enhancement
Genistein (100ug/mL)	Cooperative	10	inhibited global synthesis of DNA, RNA and proteins, leading to antimicrobial activity (16328542)	Antimicrobial		
			stabilized covalent topoisomerase II-DNA cleavage complex, which may contribute to its	Antimicrobial		

			antimicrobial effect (14738897)			
Berberine (500ug/mL)	main therapeutic component	16	inhibited microbial division protein FtsZ to produce antimicrobial effect (18275156, 21060782)	antimicrobial		
			effluxed by a multidrug pump (10677479)	Efflux-mediated multidrug resistance	Orobol inhibited the multidrug pump, thereby potentiated berberine's antimicrobial activity (12952418)	bioavailability enhancement
Orobol	sensitizer			potentiation		

Supplementary Table S4.4: Targets and potency-enhancing molecular interaction modes in 5 fully sub-potent natural product combinations with potencies of a non-principal component increased by 10–100 fold.

Ingredient	Role in Combination	Dose Reduction Index	Theapeutic Effect or Response	Effect type	Synergistic Action for Effect or Against Negative Response	Type of Synergism

Vanillin (0.6mg/mL)	main therapeutic ingredient	8	inhibited CYP53A15 to produce antifungal effect (18505250)	Antifungal	(+/-)-pinoresinol caused damage to fungal plasma membrane(20657496) to enhance vanillin's transport across fungal membrane (15868144)	bioavailability enhancement
			polymerized by laccase lacA to reduce its antifungal effect	Counteractive action		
			catabolized by vanillin dehydrogenase vdh (22057861)	Counteractive action		
4-hydroxy-3-methoxycinnamaldehyde (0.4mg/mL)	Cooperative	2	antifungal mechanism unreported			
(+/-)-pinoresinol (1mg/mL)	Cooperative	10	caused damage to fungal plasma membrane to produce antifungal effect (20657496)	Antifungal		
Vanillin (0.6mg/mL)	main therapeutic	3	inhibited CYP53A15 to produce antifungal	Antifungal	Scopoletin inhibited fungal efflux pumps (15826040)	bioavailability enhancement

Appendices

	ingredient		effect (18505250)			
			polymerized by laccase lacA to reduce its antifungal effect	Counteractive action		
			catabolized by vanillin dehydrogenase vdh (22057861)	Counteractive action	Scopoletin inhibited fungal oxidation of vanillin to enhance its bioavailability (15826040)	bioavailability enhancement
4-Hydroxy-3-methoxycinnamaldehyde (0.4mg/mL)	Cooperative	4	antifungal mechanism unreported			
Scopoletin (1.5mg/mL)	Cooperative	18.8	hindered fungi survival or germination, inhibited detoxification enzymes (15826040)	Antifungal		
berberine (125ug/mL)	main therapeutic ingredient	4.2	inhibited microbial division protein FtsZ to produce antimicrobial effect (18275156, 21060782)	antimicrobial		

Appendices

			effluxed by a multidrug pump (10677479)	Efflux-mediated multidrug resistance	chrysosplenol-D inhibited the multidrug pump, thereby potentiated berberine's antimicrobial activity (12494348)	bioavailability enhancement
chrysosplenol-D (250ug/mL)	Cooperative	10	antimicrobial mechanism unreported			
berberine (125ug/mL)	main therapeutic ingredient	4.2	inhibited microbial division protein FtsZ to produce antimicrobial effect (18275156, 21060782)	antimicrobial		
			effluxed by a multidrug pump (10677479)	Efflux-mediated multidrug resistance	chrysosplenetin inhibited the multidrug pump, thereby potentiated berberine's antimicrobial activity (12494348)	bioavailability enhancement
chrysosplenetin (250ug/mL)	Cooperative	40	antimicrobial mechanism unreported			

curcumin (3.1uM)	main therapeutic ingredient	3.1	downregulated Notch1 and Bcl-xL to inactivate NFkB, thereby promoting growth inhibition and apoptosis (16628653)	Anticancer, growth inhibition, apoptosis	isoflavone inhibited Notch, NFkB and AkT, and activated P53(22200028) to complement curcumin's action on Notch1 and Bcl-xL (16628653), thereby further promoting apoptosis	Complementary action
			activated P38, thereby downregulating Bcl2, survivin and AkT signaling to promote apoptosis (19676105)	Anticancer, apoptosis	isoflavone inhibited Notch, NFkB and AkT, and activated P53(22200028) to complement curcumin's action on Bcl2, survivin and AkT (19676105), thereby further promoting apoptosis	Complementary action
			inhibited AKT-mTOR pathway to promote anticancer effect (21450334)	Anticancer, growth inhibition		
isoflavone (183uM)	Cooperative	18.3	inhibited Notch, NFkB and AkT, and activated P53 to promote apoptosis (22200028)	Anticancer, apoptosis		

Appendices

Supplementary Table 5.1: FDA endorsed mobile apps

Device Name	Applicant	510(k) Number	type	measure	disease
AIDERA DIASEND SYSTEM	AIDERA AB	K101806	data transmitter		
AIRSTRIP OB	MP4 SOLUTIONS, LP	K042082	monitoring	fetal heart tracings; maternal contraction pattern	Obstetrics/Gynecology
AIRSTRIP OB	AIRSTRIP TECHNOLOGIES, LP	K090061	monitoring	fetal heart tracings; maternal contraction pattern	Obstetrics/Gynecology
AIRSTRIP OB	AIRSTRIP TECHNOLOGIES, LP	K090269	monitoring	fetal heart tracings; maternal contraction pattern	Obstetrics/Gynecology
AIRSTRIP REMOTE PATIENT MONITORING (RPM)	AIRSTRIP TECHNOLOGIES, LP	K110503	data viewer		
AIRSTRIP REMOTE PATIENT MONITORING (RPM) REMOTE DATA VIEWING	AIRSTRIP TECHNOLOGIES, LP	K112235	data viewer		
AIRSTRIP REMOTE PATIENT MONITORING (RPM) REMOTE DATA VIEWING	AIRSTRIP TECHNOLOGIES, LP	K121871	data viewer		
AIRSTRIP REMOTE PATIENT MONITORING (RPM) REMOTE DATA VIEWING SOFTWARE, VERSION 3.1	AIRSTRIP TECHNOLOGIES, LP	K100133	data viewer		
ALIVECOR HEART MONITOR FOR IPHONE	ALIVECOR, INC.	K122356	monitoring	ECG	cardiovascular
ASTHMAPOLIS SYSTEM	RECIPROCAL LABS CORPORATION	K121609	medical aid	actuations of prescribed MDI usage	Anesthesiology
AVITA BLUETOOTH BLOOD PRESSURE MONITOR, MODEL: BPM656ZB	AVITA CORPORATION	K072137	monitoring	systolic and diastolic blood pressure; pulse rate	
AYCAN MOBILE	AYCAN	K122260	data viewer	medical images for diagnosis from CT and	

Appendices

	DIGITALSYSTEME GMBH			MRI	
BEAM BRUSH/BEAM APP	BEAM TECHNOLOGIES, LLC	K121165	monitoring	brushing usage data	tooth decay
BIOHARNESS	ZEPHYR TECHNOLOGY CORPORATION	K113045	monitoring	ECG	cardiovascular
BODYGUARDIAN SYSTEM BODYGUARDIAN CONTROL UNIT BODYGUARDIAN CONNECT	PREVENTICE, INC.	K121197	monitoring	ECG;activity;heart rate; respiration rate	cardiovascular
CARESTREAM PACS	CARESTREAM HEALTH, INC.	K110919	data viewer	3D image	radiology
CG-5108 ACT-3L CONTINUOUS ECG MONITOR AND ARRHYTHMIA DETECTOR	CARD GUARD SCIENTIFIC SURVIVAL LTD.	K110499	monitoring	ECG	cardiovascular
CG-6108 ACT-3L CONTINUOUS ECG MONITOR & ARRHYTHMIA DETECTOR	CARD GUARD SCIENTIFIC SURVIVAL, LTD.	K081257	monitoring	ECG	cardiovascular
CG-6108 ACT-IL CONTINUOUS ECG MONITOR AND ARRHYTHMIA DETECTOR	CARD GUARD SCIENTIFIC SURVIVAL, LTD.	K101639	monitoring	ECG	cardiovascular
CG-6108 ARRHYTHMIA ECG EVENT RECORDER	CARD GUARD SCIENTIFIC SURVIVAL, LTD.	K060911	monitoring	ECG	cardiac arrhythmia
CG-6108 CONTINUOUS ECG MONITOR AND ARRHYTHMIA DETECTOR	CARD GUARD SCIENTIFIC SURVIVAL, LTD.	K071995	monitoring	ECG	

Appendices

CONFIDANT 2.5	CONFIDANT INC.	K072698	data transmitter		
CUSTOMIZED SOUND THERAPY (CST)	TINNITUS OTOSOUND PRODUCTS, LLC	K070599	treatment		Tinnitus
DASH KNEE	BRAINLAB AG	K102251	medical aid		
DATEX-OHMEDA S/5 WEB VIEWER, DATEX-OHMEDA S/5 POCKET VIEWER AND DATEX-OHMEDA S/5 CELLULAR VIEWER WITH L-WEB04 SOFTWARE	GE HEALTHCARE	K052975	data viewer	real-time patient information	
DIABETESMANAGER SYSTEM, DIABETESMANAGER-RX SYSTEM MODEL VERSION 1.1	WELLDON, INC	K100066	monitoring	glucose	diabetes
FREESTYLE TRACKER DIABETES MANAGEMENT SYSTEM	ABBOTT DIABETES CARE INC.	K020866	monitoring	glucose	diabetes
FREESTYLE TRACKER DIABETES MANAGEMENT SYSTEM	ABBOTT DIABETES CARE INC.	K020866	monitoring	glucose	diabetes
FULLY AUTOMATIC ELECTRONIC BLOOD PRESSURE MONITOR MODEL KD-931	ANDON HEALTH CO.,LTD	K102939	monitoring	blood pressure	cardiovascular
FULLY AUTOMATIC WIRELESS BLOOD PRESSURE WRIST MONITOR	ANDON HEALTH CO., LTD	K121470	monitoring	blood pressure	cardiovascular
GLUCOPHONE BLOOD GLUCOSE TEST SYSTEM, MODEL IGM-0025	INFOPIA CO., LTD	K091168	monitoring	glucose	diabetes
IBGSTAR BLOOD GLUCOSE MONITORING SYSTEM, IBGSTAR DIABETES MANAGER APPLICATION, REV D	AGAMATRIX INC	K103544	monitoring	glucose	diabetes
IGLUCOSE SYSTEM	POSITIVEID CORPORATION	K111932	monitoring	glucose	diabetes
IMCO-STAT	IMCO TECHNOLOGIES	K063392	data viewer		

Appendices

INTUITION	TERARECON, INC.	K121916	data viewer	EBT, CT, PET or MRI image	
KD-936 FULLY AUTOMATIC WIRELESS BLOOD PRESSURE MONITOR	ANDON HEALTH CO.,LTD	K120672	monitoring	blood pressure	cardiovascular
MEDAPPS REMOTE PATIENT MONITORING, MODEL MA 100	MEDAPPS, INC.	K062377	data transmitter		
MEDICALGORITHMICS REAL-TIME ECG MONITOR AND ARRHYTHMIA DETECTOR, MODEL POCKETECG	MEDICALGORITHMICS SP Z.O.O.	K090037	monitoring	heart beat, rhythm abnormalities	cardiovascular
MOBILE MIM	MIM SOFTWARE INC.	K103785	data viewer	SPECT, PET, CT, and MRI	
MOBILE MIM	MIM SOFTWARE INC.	K112930	data viewer	SPECT, PET, CT, MRI, X-ray and Ultrasound	
MOBILECT VIEWER	NEPHOSITY, INC.	K123082	data viewer	CT, MRI, X-Ray images	
MOBILE-PATIENT VIEWER	DATA CRITICAL CORPORATION	K011436	data viewer		
MOBIUS ULTRASOUND IMAGING SYSTEM	MOBISANTE, INC.	K102153	imaging		
MODIFICATION TO: CG-6108 ACT-3L CONTINUOUS ECG MONITOR AND ARRHYTHMIA DETECTOR, MODEL FG-00084	CARD GUARD SCIENTIFIC SURVIVAL, LTD.	K101703	monitoring	ECG	cardiovascular
MODIFICATION TO: POCKETVIEW ECG SOFTWARE	MICROMEDICAL INDUSTRIES, LTD.	K013311	data viewer	ECG	cardiovascular
MYGLUCOHEALTH GLUCOSE MONITORING SYSTEMS	ENTRA HEALTH SYSTEMS, LTD.	K081703	monitoring	glucose	diabetes
MYVISIONTRACK(TM)	VITAL ART AND SCIENCE INCORPORATED	K121738	monitoring	central 3 degrees metamorphopsia (visual distortion)	maculopathy
ORTHOSIZE	ORTHOSIZE LLC	K120115	medical aid		preoperative planning of orthopedic surgery

Appendices

PANOPTIC	WELCH ALLYN, INC.	K121405	imaging		
PILL PHONE	VOCEL	K060298	medical aid drug compliance		
PIXEL APP	GAUSS SURGICAL, INC.	K120473	medical aid		surgery
PIXEL APP	GAUSS SURGICAL, INC.	K121274	medical aid		surgery
PROTEUS INGESTION CONFINMATION SYSTEMS	PROTEUS BIOMEDICAL, INC.	K113070	monitoring	physiological and behavioral metrics including heart rate, activity, body angle and time-stamped user-logged events	general
REKA E100	REKA PTE LTD	K111438	monitoring	ECG	cardiovascular
RESOLUTIONMD MOBILE 3.1 MODEL RMD-MOB-31	CALGARY SCIENTIFIC, INC.	K123186	data viewer	CT and MR medical images	
RESOLUTIONMD MOBILE MODEL RMB-MOB-2X	CALGARY SCIENTIFIC, INC.	K111346	data viewer	CT and MR medical images	
RHYTHMSTAT XL	DATA CRITICAL CORP.	K971650	diagnostic	ECG	cardiovascular
SD360 DIGITAL RECORDER/SD360 HOLTER DIGITAL RECORDER	NORTHEAST MONITORING, INC.	K041901	monitoring	heart beat	cardiovascular
SILHOUETTE, MODEL 1000.01	ARANZ MEDICAL LIMITED	K070426	monitoring	external wounds	external wounds
SMARTHEART	SHL TELEMEDICINE INTERNATIONAL LTD.	K113514	monitoring	12 lead EGG and rhythm strip	cardiovascular
SPECTRUM AND SPECTRUM WITH MASTER DRUG LIBRARY	SIGMA INTL.	K042121	medical aid administration		
SURGICASE CONNECT	MATERIALISE N.V.	K113599	data transmitter	CT and MR medical images	cardiovascular
SYMCARE DIABETES MANAGEMENT PROGRAM	SYMCARE	K083263	data transmitter	glucose	diabetes

Appendices

	PERSONALIZED HEALTH SOLUTIONS, INC				
TM2005 PERSONAL MEDICAL PHONE CENTER	CARD GUARD SCIENTIFIC SURVIVAL, LTD.	K024365	data viewer	ECG, and other patient related data, (such as demographics, doctors, medical history and status, diagnoses, etc.) .	cardiovascular
VEO MULTIGAS MONITOR FOR POCKET PC, MODEL 400221	WEISSBURG ASSOCIATES	K051857	monitoring	carbon dioxide; oxygen	Anesthesiology
VESTIBULAR ANALYSIS APPARATUS	CAPACITY SPORTS, LLC	K121590	monitoring	balance	
WAVESENSE DIABETES MANAGER MODEL VERSION 1.3.4	AGAMATRIX	K101597	data transmitter	glucose	diabetes
WEB VIEWER, POCKET VIEWER AND CELLULAR VIEWER WITH L- WEB05 SOFTWARE	GE HEALTHCARE	K061994	data viewer		
WELLDON DIABETES MANAGER SYSTEM AND DIABETES MANAGER RX SYSTEM	WELLDON, INC	K112370	monitoring	glucose	diabetes
WELLDON DIABETES MANAGER SYSTEM AND DIABETES MANAGER RX SYSTEM	WELLDON, INC	K120314	monitoring	glucose	diabetes
WITHINGS BLOOD PRESSURE MONITOR	WITHINGS	K110872	monitoring	blood pressure	cardiovascular
WITHINGS, SMART BODY SCALE	ZHONGSHAN TRANSTEK ELECTRONICS CO., LTD.	K121971	monitoring	weight, BMI, body fat	