MULTI-TARGET SELECTION AND HIGH THROUGHPUT QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP MODEL DEVELOPMENT

LIU XIN

(*B.Eng, Tongji University*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2012

ACKNOWLEDGEMENTS

I would like to acknowledge and extend my heartfelt gratitude to the following persons. This dissertation would have never been possible without their constant support and aid.

First and foremost, I would like to express my heartfelt appreciation and thanks to Prof. Chen Yu Zong who has been a great mentor throughout my four and a half years' studying and research in NUS. His enthusiasm and dedication to research, his insight in science discovery, his critical thinking, his hard working spirit, and his humbleness has always been enlightening to me. In addition, his concerns about my future career and his willingness to share with me his personal experience and insightful understanding of life will always be valuable treasure throughout the rest of my life. I would like to express my utmost gratefulness to Prof Chen Yu Zong and wish him and his beloved family the very best with work and life.

Secondly, I do really appreciate Prof. Tan Tin Wee for offering me the job working as his Teaching Assistant, which has been, as he said, a "miserable" but finally turn out to be a delightful journey. I hated him for all the random odd ideas, for all the nonsenses during every tedious meeting and for keeping me and Lizhen more than busy. But now when I am reaching the end of this journey, I just realized how much I love teaching and how much more I got as return from this job, and I have already started to miss those days that I have ever spent in class with those students.

My many thanks also go to all the previous and current BIDD group members. In particularly, I would like to thank Dr. Zhang Hailei, Dr. Wang Rong, Dr. Liu Xianghui, Dr. Jia Jia, Mr. Tao Lin, for all the collaboration and the valuable friendship. My special thankfulness goes to Dr. Ma Xiaohua who treats me as a family, to Dr. Zhu Feng for the every single day we ever spent together in the same office and to Dr. Shi Zhe for all the lectures, exams and all the happiness and sadness we have be through together during the past four years. I would also like to thank my juniors, Ms. Wei Xiaona, Mr. Zhang Jingxian, Mr. Han Bucong, Ms. Qin Chu and Mr. Zhang Cheng for their assistance in my research work.

My life in Singapore would never be so cheerful without the close friendship from those fun and lovely individuals. To name a few, I would like to thank Ms. Bai Yang for her accompany from thousands of miles away throughout all these 12 years. I would also like to thank Mr. Dong Xuanchun for including me in whatever good or bad things in his life and all the jokes and bickers between us. My gratitude also goes to Ms. Liu Xinyi, Ms. Sun Jing, Ms. Du Yun, Ms. Cao Pu, Ms. Sit Wing Yee, Mr. Tu Weimin and Mr. Guo Yangfan for them being such great friends.

Last but not least, my utmost gratefulness goes to my wonderful parents and families for their everlasting love and support. I could never thank my parents more for their love for me and for them raising me up as a strong and decent young woman. I would also like to thank my newly married husband, Mr. Li Nan, for him being supportive and understanding throughout the whole time even when I have never done a single thing for him as a wife. To my beloved parents, I dedicate this thesis; on his 31st birthday, to my beloved husband, I dedicate my heart and soul forever.

Liu Xin

10 April 2012

TABLE OF CONTENTS

SUMMARY

Drugs designed to act against individual molecular targets cannot usually combat multigenic diseases such as cancers in which alternative or compensatory pathways are often activated. Thus selection of proper multi-target combinations and prediction of new molecules against these selected multiple targets are highly useful for discovering drugs with improved therapeutic efficacies by collective regulations of primary therapeutic targets, compensatory signaling and drug resistance mechanisms.

Cross-talk between pathways plays important regulatory roles in biological processes, disease processes, and therapeutic responses. Knowledge of these cross-talks is highly useful for facilitating systems level analysis of diseases, biological processes and the mechanisms of multi-targeting drugs and drug combinations. However, to our best knowledge, currently no such database exists providing this kind of information. In this work, a Pathway Cross-talk Database (PCD) is developed providing information about experimentally discovered cross-talks between pathways and their relevance to diseases and biological processes thus facilitating multi-target selection. Based on some entries stored in PCD, four combinations of anticancer kinase targets, EGFR-VEGFR, EGFR-Src, EGFR-PDGFR and EGFR-FGFR were selected as illustration and for further study.

In silico methods have been extensively explored for the discovery of multi-target drugs. Apart from drug lead optimization, predictive quantitative structure-activity relationship (QSAR) models with well-defined applicability domains (ADs) have shown promising capability in virtual screening (VS) large chemical databases for novel drug hits. Despite the good hit rates and activity assessment these QSAR models can achieve, however, these models cannot find highly novel actives outside similarity-based ADs. One possible reason is that ADs may only contain limited spectrum of active compounds. Another possible reason lies in the limited scaffold

hopping ability of the molecular descriptors, i.e. the chosen molecular descriptors may not be able to fully represent and identify molecules with similar properties yet different or novel scaffolds. Thus, an extended QSAR approach is needed aimed at finding highly novel inhibitors without compromising hit rates within similarity-based ADs. In this work, new MLR QSAR models are constructed via chemspace-wide activity regression and tested on DHFR, ACE and Cox2 inhibitors, and further applied for searching for dual inhibitors of the four combinations of anticancer kinase targets, EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src. The results show our consensus SVR QSAR models yield equivalent predictive accuracy for newly discovered chemicals and improved hit-rates and enrichment factors in identifying inhibitors from large chemical databases. In particular, our method also shows some level of capability in the identification and activity assessment of highly novel inhibitors outside similarity-based ADs.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACE	Angiotensin converting enzyme
AD	Applicability domain
AE	Adverse effect
BTFAP	Bayesian-based target-family activity profiling
CA	Carbonic anhydrase
CDK	Cyclin-dependent kinase
CoMFA	Comparative molecular field analysis
Cox2	Cyclooxygenase-2
CV	Cross validation
DBMS	Development of Database Management System
DHFR	Dihydrofolate reductase
EGFR	Epidermal growth factor receptor
ER	Estrogen receptor
FAK	Focal adhesion kinase
GGTase-I	Geranylgeranyltransferase type I
HDAC	Histone deacetylase
IGF	Insulin-like growth factor
IRS	Insulin receptor substrate
KKT	Karush-Kuhn-Tucker
kNN	*k*-nearest neighbor
mAb	Monoclonal antibody
MDDR	MDL Drug Data Report
ML	Machine learning
MLR	Machine learning regression

NCI	National Cancer Institute
NK1	Neurokinin 1
NSCLC	Non-small cell lung cancer
OODB	Object-oriented database
OOPL	Object-oriented programming language
PCD	Pathway Cross-talk Database
QSAR	Quantitative structure-activity relationship
SA-PLS	Simulated annealing-partial least squares
SVM	Support vector machine
SVR	Support vector regression
TKI	Tyrosine kinase inhibitor
VEGFR	Vascular endothelial growth factor receptor
VS	Virtual screening

LIST OF PUBLICATIONS

1. In silico prediction of adverse drug reactions and toxicities based on structural, biological and clinical data. X. Liu, Z. She, Y. Xue, Z.R. Li, S.Y. Yang and Y.Z. Chen. *Current Drug Safety*. Jul 1;7(3):225-37 (2012).
2. The Therapeutic Target Database: an internet resource for the primary targets of approved, clinical trial and experimental drugs. **X. Liu**, F. Zhu, X.H. Ma, L, Tao, J.X. Zhang, S.Y. Yang, Y.C. Wei and Y.Z. Chen. *Expert Opin Ther Targets.* 15(8):903-12 (2011).
3. Virtual screening methods as tools for drug lead discovery from large chemical libraries. X.H. Ma, F. Zhu, X. Liu, Z. Shi, J.X. Zhang, S.Y. Yang, Y.Q. Wei and Y.Z. Chen. *Curr Med Chem*. Epub ahead of print (2012).
4. 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*. Accepted (2012).
5. 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).
6. 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).
7. Trends in the Exploration of Anticancer Targets and Strategies in Enhancing the Efficacy of Drug Targeting. F. Zhu, C.J. Zheng, L.Y. Han, B. Xie, J. Jia, **X. Liu**, M.T. Tammi, S.Y. Yang, Y.Q. Wei and Y.Z. Chen. *Curr Mol Pharmacol.* 1(3):213-232 (2008).

CHAPTER 1 Introduction

Drugs designed to act against individual molecular targets cannot usually combat multigenic diseases such as cancers in which alternative or compensatory pathways are often activated. Thus prediction of new molecules against selected multiple targets is highly useful for discovering multi-target drugs with improved therapeutic efficacies by collective regulations of primary therapeutic targets, compensatory signaling and drug resistance mechanisms. In this chapter, in **Section 1.1**, the rationale of adopting multi-targeted therapy for cancers over single-targeted treatments is summarized; in **Section 1.2**, recent progresses in exploration of *in silico* methods, especially Quantitative Structure-Activity Relationship (QSAR) methods (**Section 1.3**), for the discovery of multi-targeting drugs are described.

1.1 From single- to multi-targeted cancer therapy

Due to the complex mechanisms and signaling networks involved in oncogenesis, tumor invasion and proliferation, traditional monotherapies for cancers sometimes exhibit modest effects and some patients responding to certain therapeutic agents may eventually develop drug resistance. Multi-targeting agents represent the prospect for the future targeted cancer therapies. In this section, the rationale for the multi-targeted cancer therapy is described followed by the necessity of the involvement at the system level of the complex oncogenic pathways in multi-target selection.

1.1.1 From single- to multi-targeted cancer therapy

The main challenge of clinical cancer research is to find a therapeutic approach that specifically kills malignant cells with minimum possible adverse effects (AEs).[1] However, until recently, the traditional treatment of cancers has majorly relied on cytotoxic chemotherapy.[1, 2] Recent progress in understanding the mechanisms involved in malignant transformation has offered targeted therapy,[3] i.e. compounds inhibit specific tumor targets which significantly reduce undesired AEs on normal tissues, to achieve more effective and rational cancer treatment. Though a number of agents including monoclonal antibodies (mAbs) and small-molecule tyrosine kinase inhibitors (TKIs) have been approved for clinical use or in various stages of clinical development for monotherapy of cancers, the effectiveness of these agents seem to be moderate or be reduced with the development of drug resistance. This may be partially attributed to the existence of feedback loops or the activation of alterative oncogenic pathways.[1, 2, 4, 5] For instance, targeted inhibition of epidermal growth factor receptor (EGFR) has been clinically validated in several solid tumors with a number of approved drugs.[2] EGFR and vascular endothelial growth factor receptor (VEGFR) signaling pathways are independent yet interrelated with each other.[6] EGF induces VEGF expression via activation of EGFR in human cancer cells,[6-8] and conversely, VEGF expression may decrease via inhibition of EGFR signaling pathway.[8, 9] However, it has been shown that the VEGF up-regulation independent of EGFR signaling may contribute to resistance to EGFR inhibition.[6, 10] One proposed explanation involves cyclin D1 and Bcl-xL which have been found to be overexpressed in some tumor cells.[10] Cyclin D1 associates with cyclin-dependent kinase (CDK) 4 and facilitates cell cycle progression from G1 into the S phase. Bcl-xL functions as a repressor of cell death. Both cyclin D1 and Bcl-xL expression has been shown to be positively regulated by EGFR signaling and that down-regulation of these molecules by inhibiting EGFR is believed to be critical in their proapoptotic and growth-inhibitory effects.[11-13] Additionally, it has been shown that cyclin D1 overexpression may result in increased VEGF levels.[14] High expression levels of Bcl-xL are also found to be independent of EGFR signaling,[10] which suggests a possible involvement of this antiapoptotic molecule in the resistant phenotype.

With the approval by FDA of more multi-targeting drugs such as Sorafinib and Sunitinib, discovering molecules simultaneously interfering with multiple therapeutic targets or oncogenic pathways might offer more effective clinical benefits and present the next generation of targeted therapies for cancers[1, 2].

1.1.2 Multi-target molecular scaffolds

Drugs typically interact with multiple proteins, and those interacting with selected combination of targets have found useful therapeutic applications.[15] Multi-target drugs active against selected multiple targets of the same diseases have been increasingly explored[16, 17] for achieving enhanced therapeutic efficacies and reduced drug resistance activities by simultaneously modulating a primary therapeutic target and drug response and resistance mechanisms.[18, 19] **Table 1.1** provides 32 approved and clinical trial multi-target drugs against the same diseases.[20]

Table 1.1 Literature reported multi-target drugs, targeted diseases, potencies against individual targets and cell-lines, and multi-target mode of action

Drug	Targeted Disease	Multi-targets and potency against each individual target (IC_{50}, Ki, EC_{50})	Potency against specific cell line	Multi-target mode of action
ABT-263	Advanced small cell lung cancer; Relapsed or refractory chronic lymphocytic leukemia; Relapsed or refractory lymphoid malignancies[21]	Bcl-2: <1nM Bcl-xL: <0.5nM Bcl-W: <1nM[22]	CCRF-CEM: 450nM CHLA-136: 2170nM CHLA-258: 780nM CHLA-266: 1140nM COG-LL-317: 570nM Kasumi-1: 90nM MOLT-4: 260nM NALM-6: 1080nM NB-1643: 500nM NB-EBc1: 1910nM Rh18: 200nM Rh41: 190nM RS4;11: 50nM[23]	Inhibiting Bcl-2 protein family members that regulate apoptosis and impact tumor formation, progression and chemoresistance
Afatinib	NSCLC[21]	EGFR: 0.5nM HER2: 14nM[24]	HCC827: <1nM PC9: <1nM[25]	Inhibiting tyrosine kinase receptor ERBB family members that regulate proliferation and survival at different upstream points, and act as back-up alternative for each other
AT9283	Adult solid tumors; NHL; AML; ALL; CML; MDS; Myelofibrosis[21]	AURKA: 3nM AURKB: 3nM[26]	A2780: 7.7nM A549: 12nM HCT116: 13nM HT-29: 11nM MCF7: 20nM MIA-Pa-Ca-2: 7.8nM SW620: 14nM[27]	Inhibiting Aurora kinases that regulate prophase of mitosis (Aurora A) and the attachment of the mitotic spindle to the centromere (Aurora B)

Axitinib	Metastatic pancreatic cancer; RCC; NSCLC; Breast cancer; Melanoma[28]	CSF-1: 73nM PDGFR: 1.6-5nM VEGFR2: 0.2nM[29]	HUVEC: 573nM IGR-NB8: 849nM SH-SY5Y: 274nM[30]	Inhibiting cytokine and tyrosine kinases receptors that regulate cell proliferation at different upstream points (CSF-1, PDGFR) and angiogenesis (VEGFR2)
AZD0530	Haematological malignancies; Solid tumors[28]	ABL1: 30nM SRC: 2.7nM[31]	LS180: 500nM H508: 500nM LS174T: 500nM[32] 1483: 1000nM UM-22B: 1000nM PCI-15B: 1300nM PCI-37B: 1000nM Cal-33: 600nM[33]	Inhibiting tyrosine kinases that regulate cell proliferation at different upstream points
Batimastat	Various cancers[21]	MMP-1: 5nM MMP-2: 4nM MMP-7: 6nM[34]	MDA435ILCC6: >5000nM[35]	Inhibiting MMP proteases that regulate cell invasion and proliferation (MMP-1 and 7), invasion and metastasis (MMP-2)
BMS-599626	Various cancers[28]	EGFR: 22nM HER2: 32nM[36]	AU565: 630nM BT474: 310nM GEO: 900nM HCC1419: 750nM HCC1954: 340nM HCC202: 940nM KPL-4: 380nM MDA-MB-175: 840nM N87: 450nM PC9: 340nM Sal2: 240nM ZR-75-30: 510nM[36]	Inhibiting tyrosine kinase receptor ERBB family members that regulate proliferation and survival at different upstream points
Bosutinib	CML; Leukemia; Various cancers[28]	ABL1: 1nM SRC: 1.2nM[37]	MDA-MB-435s: 9000nM Hs578T: 5900nM[38]	Inhibiting tyrosine kinases that regulate cell proliferation at different upstream points

Bupropion	Depression[21]	NET: 1900nM[39] SERT: 22000nM [40]	TE671/RD: 10500nM SH-SY5Y: 1514nM[41]	Inhibiting monoamine transporter family members that perform complementary and compensatory actions on neural activities in synapse
HKI-272	NSCL; Breast cancer; Various cancers[28]	EGFR: 92nM HER2: 59nM[42]	3T3: 700nM SK-Br-3: 2nM BT 474: 2nM A431: 81nM MDA-MB-435: 960nM SW620: 690nM[42]	Inhibiting tyrosine kinase receptor ERBB family members that regulate proliferation and survival at different upstream points
Imatinib	CML; GIST; Intestinal cancer; Myeloid leukemia; Glioma; Lung, prostate, solid tumors[28]	ABL1: 38nM[43] KIT: 100nM[44] PDGFR: 300nM[43]	BV173: 240nM EM3: 100nM K562: 560nM LAMA84: 320nM[45]	Inhibiting tyrosine kinases that regulate proliferation at different upstream points
Lapatinib	Refractory metastatic breast cancer; RCC; Bladder, head & neck, NSCLC, brain cancer[28]	EGFR: 10.8nM HER2: 9.2nM[46]	BT474: 100nM MCF-7: 4000nM T47D: 3000nM[46]	Inhibiting tyrosine kinase receptor ERBB family members that regulate proliferation and survival at different upstream points, and act as back-up alternative for each other
Midostaurin	Colon, breast, CLL, AML, GIST, solid tumors; Non-Hodgkin's lymphoma[28]	FLT3: 528nM PKC: 22nM[47]	MCF-7: 97nM[48] Canine mastocytoma cell line C2: 157nM HMC-1.1 (lacking KIT D816V): 191nM HMC-1.2 (possessing KIT D816V): 196nM[49] HEL 92.1.7: 500nM K562: 250nM[50]	Inhibiting tyrosine kinases that regulate cell proliferation at different upstream points

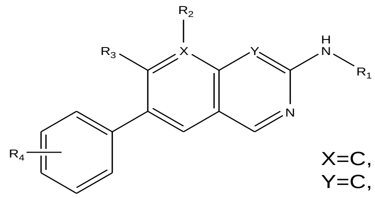
MK-5108	Various cancers[21]	AURKA: 0.064nM AURKB: 14.1nM[51]	AU565: 450nM CAL85-1: 740nM Colo205: 500nM ES-2: 1100nM HCC1143: 420nM HCC1806: 560nM HCC1954: 910nM HCT116: 270nM HeLa-S3: 2100nM MB157: 810nM MCF-7: 520nM MIAPaCa-2: 6400nM SKOV-3: 1100nM SW48: 160nM[51]	Inhibiting Aurora kinases that regulate prophase of mitosis (Aurora A) and the attachment of the mitotic spindle to the centromere (Aurora B)
Motesanib	GIST; Metastatic thyroid cancer; NSCLC; Breast, colorectal cancer[28]	KIT: 8nM PDGFR: 84nM VEGFR2: 3nM[52]	MCF-7 : >3000nM MDA-MB-231: >3000nM[53]	Inhibiting tyrosine kinase receptors that regulate proliferation (PDGFR), angiogenesis (VEGFR2), and kinase expression (KIT)
Nilotinib	ALL; CML; GIST; Leukemia[28]	ABL1: 20-60nM KIT: 27nM PDGFR: 71nM[54]	Canine mastocytoma cell line C2: 55nM HMC-1.1 (lacking KIT D816V): 10nM HMC-1.2 (possessing KIT D816V): 2363nM[49]	Inhibiting tyrosine kinases that regulate tumor growth and proliferation at different upstream points
OSI-930	Various cancers[21]	KIT: 80nM VEGFR2: 9nM[55]	H526: 9.6nM HMC-1: 9.5nM HUVEC: 10.1nM NIH-3T3: 51.5nM[56]	Inhibiting tyrosine kinase receptors that regulate cell proliferation (KIT) and angiogenesis (VEGFR2)
P276-00	Multiple myeloma; Mantle cell lymphoma; Head & neck cancers; Cyclin D1-positive melanoma[21]	CDK1: 79nM CDK4: 63nM CDK9: 20nM[57]	U266B1: 500nM RPMI-8226: 900nM[58]	Inhibiting CDK family members that are involved in cell cycle regulation (CDK1 and 4) and transcription (CDK9)

Pasireotide	Neuroendocrine tumor; Carcinoid tumor; Pancreatic neuroendocrine tumor; Pancreatic cancer[21]	SS1R: 9.3nM SS2R: 1nM SS3R: 1.5nM SS5R: 0.16nM[59]	HUVEC: 1000-10000nM[60]	Binding to multiple somatostatin receptor subtypes (i.e. 1, 2, 3, and 5) to mimic the action of natural somatostatin
Pazopanib	Advanced/metastatic renal cancer; Solid tumors; NSCLC[28]	KIT: 74nM PDGFR: 71-84nM VEGFR2: 30nM[61]	HUVEC: 21.3nM[62]	Inhibiting tyrosine kinase receptors that regulate cell proliferation and angiogenesis at different upstream points
PF-03814735	Advanced solid tumors[21]	AURKA: 5nM AURKB: 0.8nM[63]	A549: 90nM C6: 93nM H125: 150nM HCT-116: 70nM HL60: 110nM L1210: 140nM MDCK: 42nM[63]	Inhibiting Aurora kinases that regulate prophase of mitosis (Aurora A) and the attachment of the mitotic spindle to the centromere (Aurora B)
PHA-739358	CML; MHRPC[21]	AURKA: 13nM AURKB: 79nM[64]	DU145: 220nM K562: 260nM PC-3: 120nM[64]	Inhibiting Aurora kinases that regulate prophase of mitosis (Aurora A) and the attachment of the mitotic spindle to the centromere (Aurora B)
SNS-032	B-lymphoid malignancies; Advanced solid tumors[21]	CDK2: 38nM CDK7: 62nM CDK9: 4nM[65]	HCT116: <300nM[66]	Inhibiting CDK family members that are involved in cell cycle regulation (CDK2), transcription (CDK9) and CDK activating and transcription (CDK7)
Sorafenib	RCC; Hepatocellular carcinoma; NSCLC; Melanoma; Myelodyspalstic syndrome; AML; Head & neck cancer; Breast, colon, ovarian, pancreatic cancer[21]	RAF: 22nM[67] RET: 5.9nM[68] VEGFR: 20-90nM[67]	HepG2: 4500nM PLC/PRF/5: 6300nM[69] EOL-1: 0.033nM MV4-11: 0.88nM RS4;11: 12nM[70]	Inhibiting kinases that regulate angiogenesis (VEGFR2) and proliferation (BRAF), RET lysosomal degradation (RET), and Src-mediated alternative signalling (BRAF)

Sotrastaurin	Acute rejection after de novo renal transplantation[21]	PKC-alpha: 0.95nM PKC-beta: 0.64nM PKC-theta: 0.22nM[71]	PBMC: 37nM[72]	Inhibiting PKC family members that regulate the induction of transcription factors (PKC-alpha and beta) and sustainability of intracellular signals (PKC-theta) ,and in turn blocking T cell activation
SU-6668	Advanced solid tumors[21]	AURKA: 850nM AURKB: 47nM[73] FGFR: 1200nM PDGFR: 8nM VEGFR2: 2100nM[74]	H526: 8500nM[75] MO7E: 290nM[76]	Inhibiting Aurora kinases that regulate prophase of mitosis (Aurora A) and the attachment of the mitotic spindle to the centromere (Aurora B), and tyrosine kinase receptors that regulate angiogenesis (FGFR, PDGFR and VEGFR2)
Sunitinib	RCC; GIST; Breast, neuroendocrine tomors[28]	FLT3: 50-250nM[77] KIT: 1-10nM[78] PDGFR: 2nM[79] VEGFR2: 80nM[79]	Kasumi-1: 75.7nM[80]	Inhibiting tyrosine kinase receptors that regulate angiogenesis (PDGFR, VEGFR2), proliferation (FLT3), and kinase level (KIT)
TAK165	Various cancers[28]	EGFR: >25000nM HER2: 6nM[81]	BT474: 5nM UMUC-3: 1812nM T24: 91nM DU145: 1647nM PC-3: 4620nM LN-REC4: 90nM LNCaP: 53nM[81]	Inhibiting tyrosine kinase receptor ERBB family members that regulate proliferation and survival at different upstream points
TKI258	RCC[21]	FGFR3: 8nM PDGFR: 27-210nM[61]	G384D: 550nM K650E: 90nM Y373C: 90nM[82]	Inhibiting tyrosine kinase receptors that regulate survival and growth (FLT3), and angiogenesis and tumor progression (FGFR3)
VX-680	Colorectal cancer; Hematological malignancies; Various solid tumors; Hematological cancers[28]	AURKA: 0.6nM AURKB: 18nM LCK: 520nM[83]	HL60: 15nM[83]	Inhibiting Aurora kinases that regulate prophase of mitosis (Aurora A) and the attachment of the mitotic spindle to the centromere (Aurora B)
XL880	Gastric cancer; RCC; Solid tumors[21]	MET: 0.4nM VEGFR2: 0.86nM[84]	B16F10: 21nM MDA-MB-231: 4nM PC-3: 23nM[84]	Inhibiting tyrosine kinases that regulate tumor growth (c-MET) and angiogenesis (VEGFR2)

ZK 304709	Advanced solid tumors[21]	CDK1: 50nM CDK2: 4nM CDK4: 61nM CDK7: 85nM CDK9: 5nM[85]	BON: 129nM QGP-1: 79nM[86]	Inhibiting CDK family members that are involved in cell cycle regulation (CDK1, 2 and 4), transcription (CDK9) and CDK activating and transcription (CDK7)

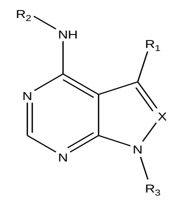
Some molecular scaffolds have been found in high percentages of multi-target agents against selected targets. For instance, the six scaffolds in **Figure 1.1** are reportedly contained in high percentages of the published dual inhibitors of tyrosine kinase pairs EGFR-PDGFR, PDGFR-Src, EGFR-Src, EGFR-FGFR, VEGFR-Lck, Src-Lck, and PDGFR-FGFR published before 2010.[87] The seven scaffolds in **Figure 1.2** are in high percentages of the published dual inhibitors of serotonin reuptake paired with noradrenaline transporter, H3 receptor, 5-HT1a receptor, 5-HT1b receptor, 5-HT2c receptor and Neurokinin 1 (NK1) receptor respectively.[88] Some scaffolds have been found to form multi-target activity scaffolds with their structural analogues having significantly different potencies against multiple targets.[89] For instance, the two scaffolds in **Figure 1.3** are in some inhibitors of carbonic anhydrase (CA) I, II and IX and some inhibitors of protein kinase B (PKB) Akt1 and Akt2, mitogen- and stress-activated protein kinase 1 (MSK1) and ribosomal S6 kinase 1 (RSK1) respectively, each with close analogues showing highly different potencies against different targets.[89] In particular, analogues a and b of scaffold A, and analogues b and c of scaffold B show markedly different pIC_{50} values (activity cliff). These and other multi-target scaffolds appear to be the backbone of multi-target inhibitors of selected targets, and specific variations of side-chain groups of these scaffolds seem to be sufficient to significantly alter multi-target activities. This suggests that structural and physicochemical properties are important for distinguishing multi-target inhibitors, which can be explored for predicting polypharmacology.[20, 87]



X=C, N
Y=C, N

Scaffold A

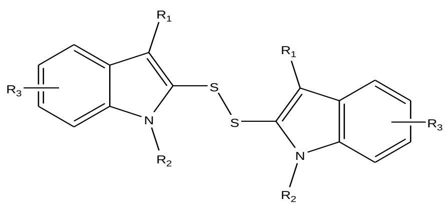
PDGFR-Src: 76.1%
PDGFR-FGFR: 67.4%
EGFR-PDGFR: 63.8%
EGFR-FGFR: 54.9%
EGFR-Src: 33.9%
VEGFR-Lck: 27.9%



X=C, N

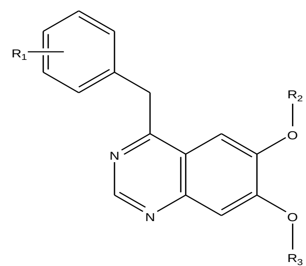
Scaffold B

Src-Lck: 57.1%
VEGFR-Lck: 29.5%
EGFR-Src: 25.9%



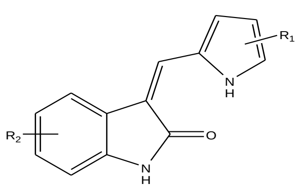
Scaffold C

EGFR-Src: 19.6%



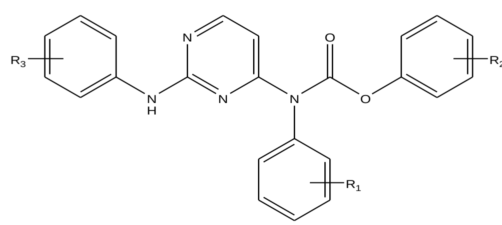
Scaffold D

EGFR-FGFR: 32.4%
EGFR-Src: 4.5%
Src-Lck: 1.8%
EGFR-PDGFR: 1.7%



Scaffold E

PDGFR-FGFR: 17.8%
EGFR-PDGFR: 8.6%
EGFR-FGFR: 7.0%
PDGFR-Src: 6.9%
EGFR-Src: 2.7%
Src-Lck: 1.8%



Scaffold F

Src-Lck: 37.5%
VEGFR-Lck: 34.4%

Figure 1.1 Six scaffolds contained in high percentages of the dual inhibitors of tyrosine kinase pairs.

These tyrosine kinase pairs include EGFR-PDGFR, PDGFR-Src, EGFR-Src, EGFR-FGFR, VEGFR-Lck, Src-Lck, PDGFR-FGFR, and PDGFR-Src published before 2010. The percentage value behind each target-pair indicates the percentage of known dual inhibitors of the target-pair that contain this scaffold.

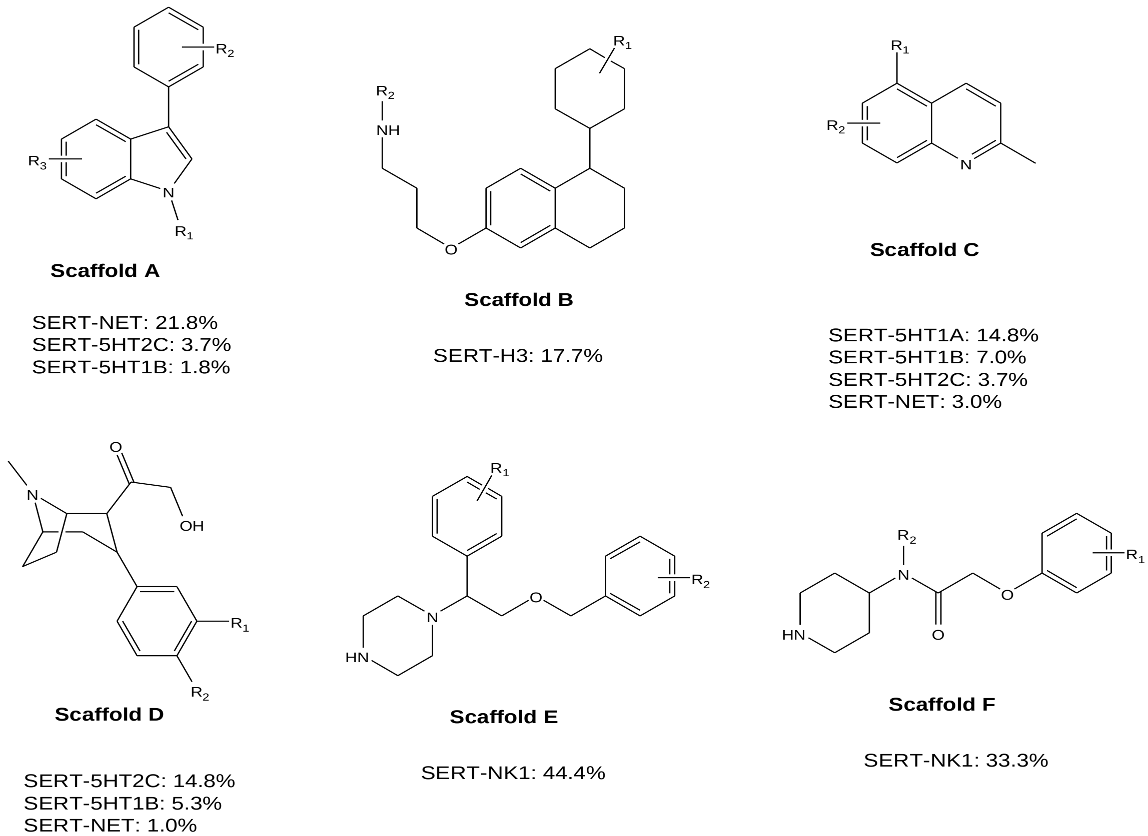


Figure 1.2 Seven scaffolds reportedly contained in high percentages of the published dual inhibitors of serotonin reuptake paired with other targets.

The listed dual inhibitors are those of serotonin reuptake paired with noradrenaline transporter, H3 receptor, 5-HT1a receptor, 5-HT1b receptor, 5-HT2c receptor, Melanocortin 4 receptor and Neurokinin 1 receptor respectively. The percentage value behind each target-pair indicates the percentage of known dual inhibitors of the target-pair that contain this scaffold.

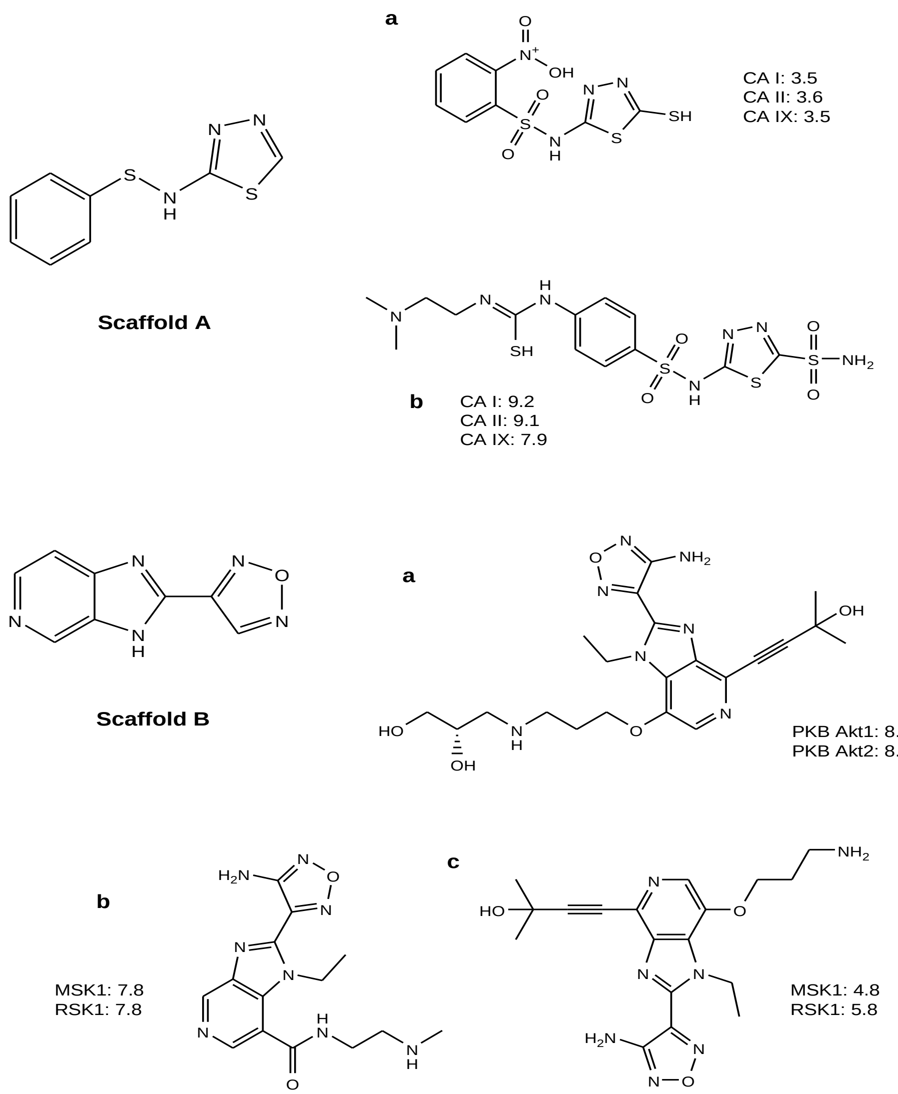


Figure 1.3 Two molecular scaffolds in some multi-target inhibitors of CAI, CAII and CAIX and some inhibitors of Akt1, Akt2, MSK1 and RSK1 respectively.

Each of these two scaffolds are with representative multi-target analogues showing potencies in pIC_{50} against respective target combinations. In particular, analogues a and b of scaffold A, and analogues b and c of scaffold B show markedly different pIC_{50} values (activity cliff).

1.1.3 Proposed prospect of multi-target selection

Modern drug discovery is primarily focused on the search or design of drug-like molecules, which selectively interact and modulate the activity of one or a few selected therapeutic targets.[16, 90, 91] One challenge in drug development is to choose and explore promising targets from a growing number of potential targets.[92] Target selection is of significant importance not only for achieving therapeutic efficacy but also for increasing drug development odds, given that few innovative targets have made it to the approved list each year (12 innovative targets in 1994–2005[93] and 10 new human targets in 2006–2010[94] for small molecule drugs).

Traditionally, the selected drug target is a single gene or gene product based on genetic analysis and biological observations.[95] Pathway analysis approaches have also been incorporated in the process of target selection[95, 96] especially for cancers due to the reliance of these signaling pathways on the action of protein kinases whose dysregulation largely contributes to oncogenesis and tumor progress.[95] However, drugs targeting specific single pathways exhibit limited efficacies, undesired AEs and resistance profiles often resulted from the multi-factorial mechanisms of cancers[95] and the activation of alterative pathways[1, 2, 4, 5] or pathway cross-talks.[97]

One example has been described in **Section 1.1.1** that the VEGF up-regulation independent of EGFR signaling may contribute to resistance to EGFR inhibition in treating non-small cell lung cancer (NSCLC).[6, 10] Another instance can be illustrated by the cross-talk between insulin-like growth factor (IGF) signaling and integrin signaling pathways that affects the phenotype of breast cancer.[97] IGFs protect breast cells from apoptosis and promote survival and IGF signaling has been proven to be a fit drug target for the treatment of breast cancer.[98, 99] Integrin signaling plays important role in the development and progression of tumors in breast cancer.[100] Moreover, the dependence of the IGF system on Integrin signaling pathway has also been demonstrated. For example, $\alpha v\beta 3$ integrin associates with IGF1R and alters IGF-1 stimulated signaling and cell migration.[101] Another mechanism of the interaction between IGF and integrin signaling pathways may recruit focal adhesion kinase (FAK) and insulin receptor substrate (IRS) proteins as mediators.[97] FAK is a primary mediator of integrin signaling.[97] The activation of IRS-1 has been shown to be associated with IGF mediated proliferation, while IRS-2 is involved in cell motility.[97] FAK has been reported to be activated by IGF1R[102] and IRS proteins are substrates of FAK.[103] Furthermore, IGF promotes the redistribution of FAK and IRS-2 to membrane terminals of breast cancer cells during cell migration.[97] Therefore, the integrin occupancy is required for the maximal effect of IGF stimulated phenotypes and the IGF system can feed into the integrin system to mediate inside-out signaling.[97] Thus, although modulating a single target has been proven to be beneficial, targeting multiple signaling pathways, especially cross-talking pathways e.g. IGF and integrin systems simultaneously to inhibit the advancement of IGF-responsive breast cancer, may prove more efficacious. [97]

Therefore, knowledge of pathway cross-talks promises to supplement and facilitate current target, especially multi-target, discovery and multi-target therapeutic strategies. Increasingly accumulated information on experimentally determined pathway cross-talks is readily available in published literature. However, to our best knowledge, no such database is available to comprehensively collect and provide such information in an organized pattern. To this end, in **Chapter 3**, a Pathway Cross-talk Database (PCD) is developed to fill in this blank thus facilitating the multi-target selection in drug discovery for achieving enhanced therapeutic efficacies and reduced drug resistance activities.

1.2 *In silico* prediction of multi-target agents

There have been increasing interests in discovering multi-target drugs[104] by means of experimental and *in silico* methods.[20, 105] In particular, a number of *in silico* methods have been used for predicting multiple targets of known drugs and newly designed molecules.[20] These methods are broadly classified into fragment-based, structure-based and ligand-based methods. Fragment-based methods combine multiple structural frameworks of active molecules of individual target into a single molecule that binds to multiple targets.[106] Structure-based methods, such as molecular docking,[107-109] target-site structural similarity[110] and receptor-based pharmacophore searching,[111] explore target site structural features to find binding molecules with structural and energetic complementarity. Ligand-based methods use such techniques as similarity searching,[112, 113] drug side effect similarity,[114] quantitative structure-activity relationships (QSAR),[115-121] and machine learning methods[87, 88] to select molecules with structural and physicochemical profiles matching those of the known active molecules. In this section, recent progresses are described in exploring these methods for predicting polypharmacology aimed at multi-target drug discovery.

1.2.1 Fragment-based methods for prediction of multi-target agents

Fragment-based approaches have also been explored for designing multi-target agents.[106] One method, framework combination, incorporates essential binding features into a single lead molecule by linking, fusing or merging the frameworks of two selective molecules.[106] However, this method may in some cases generate large, complex and less drug-like molecules.[106] Drug-likeness can be retained if the degree of framework overlap is maximized and the size of the selective ligands minimized. Another method, screening-based method, searches chemical (fragment) libraries to find multi-target fragment hits possibly with weak activities, followed by optimization of the fragment into more potent multi-target active agents.[106] Optimizing fragments with weak multiple activities into potent multi-target drug-like agents can be more easily achieved for targets sharing a conserved binding site.[122] As binding sites become more dissimilar, it remains a challenge to design agents with potent multi-target activities, *in vivo* efficacy and safety profiles. One solution is to explore synergistic targets, such that multi-target agents with modest activity against one or more of these synergetic targets may still produce similar or better *in vivo* effects compared to higher-affinity target-selective compounds.[123]

1.2.2 Structure-based methods for prediction of multi-target agents

Two structure-based methods, molecular docking and receptor-based pharmacophore searching, have been extensively used for facilitating the identification of multi-target molecules. In particular, molecular docking method does not require knowledge about known active compounds and their structural features or frameworks, but in some cases may have limited capability in account of target structural flexibility and specific chemical features of drug binding. To improve virtual screening performance, molecular dynamics enhanced molecular docking method has been used in virtual screening against the individual targets in HIV and its associated opportunistic pathogens to find multi-target agents such as KNI-764 that inhibits both HIV-1 protease and malarial plasmepsin II enzyme.[124] Molecular docking and pharmacophore matching methods have been used for identifying dual-inhibitors of two anti-inflammatory targets, PLA2 and LTA4H-h, in the arachidonic acid metabolic network.[125] Combined receptor-based pharmacophore searching and molecular docking have been used for identifying multi-target Chinese herbal ingredients against four anti-inflammatory targets cyclooxygenases 1 & 2, p38 MAP kinase, c-Jun terminal-NH2 kinase and type 4 cAMP-specific phosphodiesterase.[126]

1.2.3 Ligand-based methods for prediction of multi-target agents

Some ligand-based methods have also been used for identifying multi-target active compounds. In particular, a number of multi-target QSAR models have been developed for identifying multi-target kinase inhibitors,[115] dual action anti-Alzheimer and anti-parasitic GSK-3 inhibitors,[116, 117] HIV-HCV co-inhibitors,[118] and active agents against multiple bacterial,[119] fungal[120, 121] and viral[119] species have been developed by incorporating multi-target or species variations of binding-site features into the multi-target dependent molecular descriptors or species-dependent molecular descriptors, and stochastic Markov drug-binding process models. These multi-target QSAR models have been reported to achieve high retrieval rates of 72%~85% and moderately low false-hit rates of 15%~28%.[119-121] Development of multi-target QSAR models may be limited by the inadequate number of drug data for some of the targets or species. Moreover, the molecular size of the testing drugs needs to be in a certain range for accurate computation of multi-target dependent or species-dependent molecular descriptors, which in some cases may also affect one's capability for developing multi-target QSAR models.[121]

Another ligand-based method, machine learning method, has also been explored as virtual screening tools for multi-target drug discovery. Combinatorial SVM models for searching dual inhibitors of 11 kinase pairs have been developed, for which *in silico* tests have shown reasonably good dual kinase inhibitor yields (12.2%-57.3%), hit rates (0.22%~4.3%), and selectivity against individual kinase inhibitors (individual kinase inhibitor false selection rates 3.7%-48.1% for the same kinase pair and 0.98%-4.77% for other kinases) in screening 13.56 million compounds.[88] Some of the SVM selected virtual hits that passed drug-like filter and molecular docking have been tested in bioassays, which have found that 3 of the 19 selected dual Abl and PI3K inhibitor hits,[127] 1 of the 21 selected dual VEGFR2 and Src inhibitor hits[128] and 1 selected dual EGFR and VEGFR inhibitor hit[129] are active. Combinatorial SVM has also been applied for predicting dual target serotonin reuptake inhibitors of 7 target pairs, and *in silico* tests have shown similar level of dual target inhibitor yields (22.0%~83.3%), hit rates (0.12%~12.6%), and selectivity against individual target inhibitors (individual target inhibitor false selection rates 2.2%-29.8% for the same target pair and 0.58%-7.1% for other similar targets) in screening 17 million compounds.[88]

1.3 Predictive QSAR models as virtual screening tools

Apart from drug lead optimization, QSAR models have been developed for searching drug leads, particularly novel ones, from large chemical libraries.[130-137] These models achieve good hit rates and activity assessment by pharmacophoric-shim adjusted molecular docking (PSA-Docking),[130-132] Bayesian-based target-family activity profiling (BTFAP),[133] and machine learning regression (MLR) of known actives[134-137] within applicability domains (ADs) defined by binding-mode constraints,[130] Baysian active-inactive boundaries,[133, 138] and range-based and distance-based similarity to the known actives.[139, 140] In particular, MLR requires no knowledge of target 3D structure or target-family activity profiles.[141] A few examples of recent MLR QSAR models VS applications are highlighted below.

1.3.1 Discovery of novel D1 dopaminergic antagonists

Dopamine receptors are implicated in many neurological processes, including motivation, pleasure, cognition, memory, learning, and fine motor control, as well as modulation of neuroendocrine signaling.[142] Abnormal dopamine receptor signaling and dopaminergic nerve function is implicated in several neuropsychiatric disorders[142] and makes dopamine receptors common neurologic drug targets. Dopamine D1 receptor antagonists inhibited cell depolarization by preventing the activation of D1 receptor. However, the number of current drugs targeting D1 receptor is limited with 3 approved for marketing and another 2 under preclinical studies.[21] QSAR models were developed by comparative molecular field analysis (CoMFA), simulated annealing-partial least squares (SA-PLS), *k*-nearest neighbor (kNN), and support vector machines (SVM) approaches for 48 antagonists of the dopamine D1 receptor and applied to the VS of chemical databases to discover novel potential antagonists.[135] Validated QSAR models were used to mine 3 publicly available chemical databases: the National Cancer Institute (NCI) database, the Maybridge database and the ChemDiv database and resulted in 54 consensus hits. 5 of these 54 virtual hits were previously reported as dopamine D1 ligands, but were not included in the original dataset. A small fraction of the purported D1 ligands did not contain a catechol ring found in all known dopamine full agonist ligands, suggesting that they may be novel structural antagonist leads.[135]

1.3.2 Discovery of novel histone deacetylase (HDAC) inhibitors

Histone deacetylases (HDACs) modulate chromatin structure and transcription.[143] HDAC inhibitors have long been used in psychiatry and neurology as mood stabilizers and anti-epileptics. In more recent times, HDACs have become emerging target for the cancer treatment. In another work of Tropsha's group, QSAR models were generated by Tang et al. by kNN and SVM approaches for 59 diverse class I HDAC inhibitors.[137] Validated consensus QSAR models were then used to virtual screen 3 million compounds from 4 chemical databases: National Cancer Institute (NCI) database, Maybridge database, ChemDiv database and ZINC database. The searches resulted in 48 consensus hits, including 2 reported HDAC inhibitors that were not included in the original data set. 4 virtual hits with novel structural features were purchased and tested using the same biological assay that was employed to assess the inhibition activity of the training set compounds. 3 of these 4 compounds were confirmed active with the best inhibitory activity (IC_{50}) of 1 μM.[137]

1.3.3 Discovery of novel Geranylgeranyltransferase type I (GGTase-I) inhibitors

Geranylgeranyltransferase posttranslationally modify proteins by adding an isoprenoid lipid called a prenyl group to the carboxyl terminus of the target protein. This process, called prenylation, causes prenylated proteins to become membrane-associated due to the hydophobic nature of the prenyl group. Most prenylated proteins are involved in cellular signaling, wherein membrane association is critical for function.[144] GGTase-I inhibitors have therapeutic potential to treat inflammation, multiple sclerosis, atherosclerosis, and many other diseases.[145, 146] In a recent study, Peterson et al. constructed kNN, GA-PLS and automated lazy learning QSAR models for

48 diverse GGTase-I inhibitors and used the validated models to VS 9.5 million commercially available chemicals.[136] This yielded 47 consensus virtual hits, 7 of which were with novel scaffolds. These 7 virtual hits were further tested *in vitro* and all were found to be bona fide and selective micromolar inhibitors.[136]

Despite the good hit rates and activity assessment these models can achieve, however, these models cannot find highly novel actives outside similarity-based ADs. One possible reason is that ADs may only contain limited spectrum of active compounds. Another possible reason lies in the limited scaffold hopping ability of the molecular descriptors, i.e. the chosen molecular descriptors may not be able to fully represent and identify molecules with similar properties yet different or novel scaffolds. Thus, an extended QSAR approach is needed aimed at finding highly novel inhibitors without compromising hit rates within similarity-based ADs. In **Chapter 4,** new MLR QSAR models are constructed via chemspace-wide activity regression and tested on dihydrofolate reductase (DHFR), angiotensin converting enzyme (ACE) and cyclooxygenase-2 (Cox2) inhibitors, and further applied for VS of EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src dual inhibitors in **Chapter 5.**

1.4 Objectives and outline of this work

As described in previous sections, knowledge of pathway cross-talks is of significant importance to supplement and facilitate current multi-target discovery and therapeutic strategies. Increasingly accumulated information on experimentally determined pathway cross-talks is readily available in published literature. However, no such database is available to comprehensively collect and provide such information in an organized pattern. On the other hand, despite that the current QSAR models can achieve satisfactory hit rates and activity assessment, however, the ability of these models for yielding highly novel inhibitors are still limited, especially for those are outside similarity-based ADs. Therefore, in this work, we majorly aim to achieve the following two objectives:

1) To develop a database comprehensively collect and provide experimentally determined pathway cross-talks to facilitate the multi-target selection in drug discovery for achieving enhanced therapeutic efficacies and reduced drug resistance activities.
2) To develop an extended QSAR method via chemspace-wide activity regression that is capable of finding highly novel single- and multi-target inhibitors while without compromising hit rates within similarity-based ADs.

In summary, this dissertation is organized in the following manner:

In **Chapter 1,** the rationale of the multi-targeted cancer therapies is described coupled with the importance of employing knowledge of pathway cross-talks facilitating this process. A list of *in silico* methods, e.g. QSAR method, for the prediction of the multi-target agents is reviewed. In particular, the performance of validated QSAR models screening large chemical databases for virtual hits is also summarized.

In **Chapter 2,** details of the methods used in this work are described. In particular, the strategy for developing a Pathway Cross-talk Database is presented in every detail together with the data preparation process, the molecular descriptors calculation, mathematical models of various statistical learning methods used for the high throughput QSAR model development in this work, and the model evaluation methods.

In **Chapter 3,** a Pathway Cross-talk Database (PCD) is developed providing information about experimentally discovered cross-talks between pathways and their relevance to diseases and biological processes, mechanism of multi-target drugs and drug combinations. In this chapter, the data source, structure and access of PCD are introduced in details. The usefulness of PCD in facilitating system level studies of diseases and mechanism of drug combinations and, especially, multi-targeting drugs is also demonstrated.

In **Chapter 4,** a high throughput SVR QSAR approach is developed via chemspace-wide activity regression aimed at finding highly novel inhibitors without compromising hit rates within similarity-based applicability domains. This SVR QSAR approach is tested on DHFR, ACE and Cox2 inhibitors for predicting the activities of “new” inhibitors reported after the year of 2010 and for identifying inhibitors from large chemical databases.

4 combinations of 5 anticancer kinases, EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src, are selected in **Chapter 3** as some of the promising anti-NSCLC drug targets by the systems level analysis of the cross-talks between signalings initiated by these kinases. Thus in **Chapter 5,** the SVR QSAR approach is applied as the VS tool for searching dual inhibitors of these kinase combinations.

Finally, in the last chapter, **Chapter 6**, major findings and contributions of current work for the development and application of PCD and the high throughput SVR QSAR approach are discussed. Limitations and suggestions for future studies are also rationalized in this chapter.

CHAPTER 2 Materials and Methods

2.1 Development of systems biological network database

Database development has shown a broad spectrum of application in scientific research. Specifically, system biological databases aiming at providing comprehensive and systematic information for bioinformatics and pharmaceutics-related research have been widely utilized in the study of mechanism of diseases, identification of rational drug targets and discovery of novel drug hits, multi-targeting drugs and drug combinations and etc. Despite their various applications in biological and pharmaceutical research, the general strategy adopted for constructing these databases is similar. In this section, the basic strategy for developing knowledge-based systems biological network databases is demonstrated, which will then be extended to construct Pathway Cross-talk Database (PCD). More details on this database will be introduced later in **Chapter 3**.

Generally, the development of a database is a process including rational architecture design, information accumulation, optimal data storage and user-friendly data access and representation.

2.1.1 Rational architecture design

Before constructing any bioinformatics databases, a rational design of architecture will help us to define the scope of the database, focus on certain pharmaceutical problem, and pave the way for the information collection. At this stage, the objective and content of the database should be seriously considered. As summarized in **Chapter 1**, cross-talk between pathways plays important regulatory roles in biological processes, disease processes, and therapeutic responses. Knowledge of these cross-talks is highly useful for facilitating systems level analysis of diseases, biological processes and the mechanisms of multi-targeting drugs and drug combinations. However, currently there is no such database. Developed in the year of 2008, the Pathway Cross-talk

Database (PCD) was designed to provide information about experimentally discovered cross-talks between pathways and their relevance to diseases and biological processes, mechanism of multi-target drugs and drug combinations.

2.1.2 Information mining for system biological databases

Generally, a knowledge-based bioinformatics database is designed to provide sufficient domain knowledge on a specific subject in biology and pharmacology. Take PCD as an example, PCD was designed to provide information about experimentally discovered cross-talks between pathways thus facilitating the understanding of mechanisms of diseases and cellular processes, and discovery of multi-target drugs and drug combinations. For a single entry in PCD, knowledge is incorporated at various levels including genes, ligands, proteins, distinct single pathways and cross-talk networks.

The information planned to be integrated can be selected from a comprehensive search of literature and research publications. In light of the diversity of information types, the methods used for data collection vary, but one thing in common is to seek data from reliable resources. At present, no ready index or library is available and almost all the relevant information is scattered in the huge amount of biological and medical literature. Therefore, literature information extraction is considered to be one of the most feasible ways for information mining. It is generally agreed that literature are typically unstructured data source, and the terms used in different sources, which may be in synonymous name, various abbreviations, or totally different expression, are difficult to be recognized by automatic language processing. An automated literature information extraction system solely relying on computational recognition, thus, cannot be invented to gather information from literature both efficiently and accurately.

In this work, automatic text mining methods with manual reading process was combined. Automated text retrieval programs developed in Perl were used to screen the literature that contained the key words in the local Medline abstract packages.[147] Then, the useful subject information was picked up manually from these matched Medline abstract. If necessary, the full literature was referred to facilitate information searching. Meanwhile, in many cases, the relevant information about the same subject could also be found in the same literature. Therefore, in the first step, not only subjects but also relevant information could be obtained and recorded. In the second step, detailed biological information of subjects was automatically selected from some general or specific biological databases, such as SwissProt, KEGG and etc., by text mining program. Likewise, other information derived from the subjects was also extracted from the corresponding databases in the same way. On collecting sufficient high quality information, data storage, organization and management and design of database structure is the next step, which will be described in the next section.

2.1.3 Data organization and database structure construction

A good database system enables users to create, store, organize, and manipulate data effectively and efficiently. By integrating databases and web sites, users and clients can open up possibilities for data access and dynamic web content. An integrated information system of a database should be constructed according to some standardization strategies as follows:

1) Establishment of standardized data format and appropriate data model
2) Database structure construction
3) Development of Database Management System (DBMS)

Since the original data information collected in previous section is independent, the first major activity of a database construction process includes creation of digital files from these information fragments and construction of an appropriate data model.

2.1.3.1 The database model

Database model is an integrated collection of concepts for describing data, relationships between data, and constraint on data. In other words, a database model is a specific description on how a database is structured and used. The basic ways of constructing databases include:

1) The flat file model
2) The hierarchical model
3) The network model
4) The relational model
5) The object-oriented model

The flat-file model is the simplest data model, which is essentially a plain table of data.[148] Each item in the flat file, called a record, corresponds to a single, complete data entry. A record is made up by data elements, which is the basic building block of all data models, not just flat files. The flat-file data model is relatively simple to use; however, it is insufficient for large databases.

The hierarchical data model organizes data in a tree-like structure (**Figure 2.1**).[149] It has been used in many well-known database management systems. The structure allows representing information using parent/child relationships: each parent can have many children, but each child has only one parent (also known as a 1-to-many relationship).[149] All attributes of a specific record are listed under an entity type. This database structure was one of the first used because it lends itself very well to linear type storage mediums, such as the data tapes that were used when database were first created. However, this model has many issues that hold it back now that we require more sophisticated relationships. It requires data to be repetitively stored in many different entities. The database can be very slow when searching for information on the lower entities.[150]

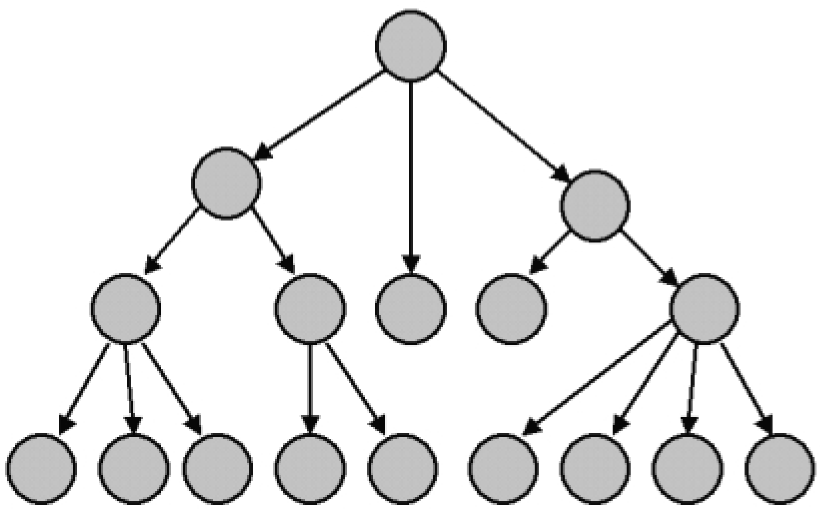


Figure 2.1 The hierarchical data model

In most cases, the relationships of data would be arbitrarily complex (**Figure 2.2**). In this model, some data are more naturally modeled with multiple parents per child. So, the network model permits the modeling of many-to-many relationships in data. This model, thus, can handle varied and complex information while remaining reasonably efficient. Even so, the biggest problem with the network data model is that databases can get excessively complicated.

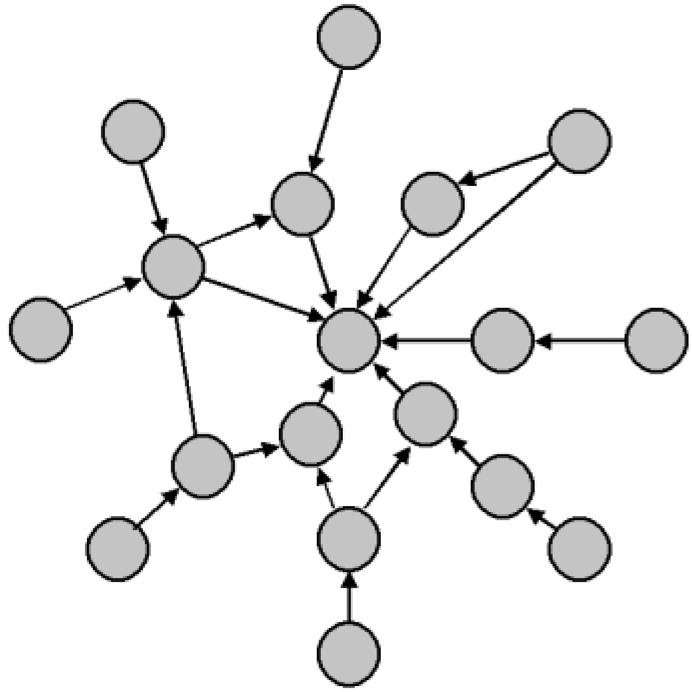


Figure 2.2 The network data model

	Data item 1	Data item 2	Data item 3	……
Record 1				
Record 2				
Record 3				
……				

Figure 2.3 The rational data model

The relational model was formally introduced in 1970[151] and has been extensively used in biological database development (**Figure 2.3**). The model is a much more versatile form of database. On the basis of this kind of data model, a novel system named relational database management system[152] is established. A relational database allows the definition of data structures, storage and retrieval operations and integrity constraints. In such a database the data and relations between them are organized in tables.

A relational database consists of multiple tables of data, related to one another by columns that are common among them.[151] Each table is a collection of records and each record in a table contains the same fields.[151] Therefore, if the database is relational, we can have different tables for different information. And the common columns, such as entry ID, can be used to relate the different tables. Relational database is the predominant form of database in use today, especially in biological research field.

The object-oriented database (OODB) paradigm[153-155] is "the combination of object-oriented programming language (OOPL) systems and persistent systems".[156] "The power of the OODB comes from the seamless treatment of both persistent data, as found in databases, and transient data, as found in executing programs".[156] The database functionality is added to object programming languages in object database management systems, which extend the semantics of the C++, Smalltalk and Java object programming languages to provide full-featured database programming capability. The combination of the application and database development with a data model and language environment is a major advantage of the object-oriented model. As a result, applications require less code, use more natural data modeling, and code bases are easier to maintain.

2.1.3.2 Construction of relational database structure

The relational model has been used in our system biological network databases. It represents relevant data in the form of two-dimension tables. Each table represents relevant data collected. The two-dimensional tables (**Figure 2.4**) for the relational database include the entry ID list table, the main information table, which contains a record for the basic information of each entry, data type table, which demonstrates the meaning represented by different number, and reference information table, which gives the general reference information following by different PubMed ID in Medline.[147]

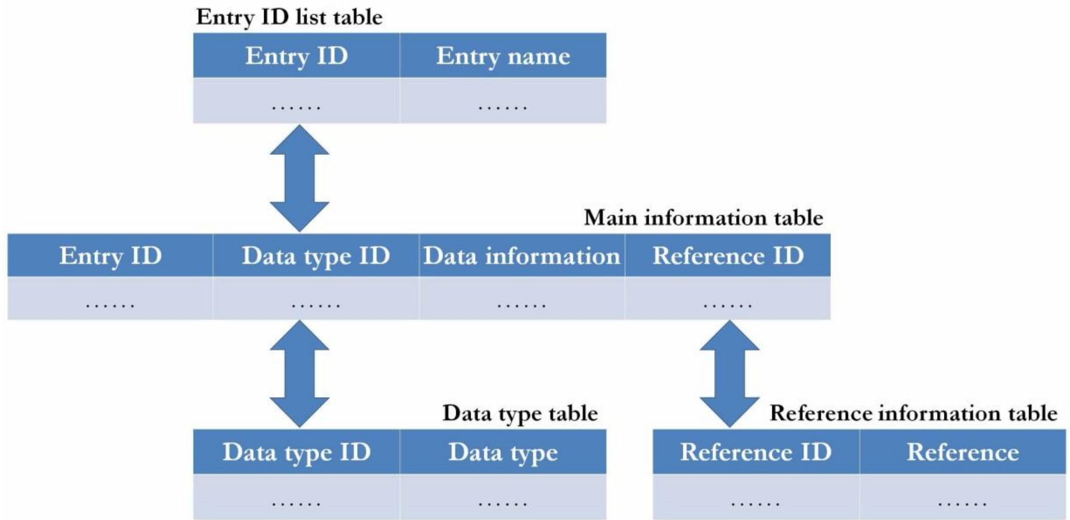


Figure 2.4 Logical view of databases

Figure 2.4 is a general logical view of databases. It shows the organization of relevant data into relational tables. In these tables, certain fields may be designated as keys, by which the separated tables can be linked together to facilitate searching specific values of that field. Commonly, in relational table, the key can be divided into two types. One is primary key, which uniquely identifies each record in the table. Here it is a normal attribute that is guaranteed to be unique, such as entry ID in entry ID list table with no more than one record per entry. The other is foreign key, which is a field in a relational table that matches the primary key column of another table. The foreign key can be used to cross-reference tables. For example, in tables of our databases, there are two foreign keys: Data type ID and Reference ID. According to **Figure 2.4**, a connection between a pair of tables is established using a foreign key. The two foreign keys make three tables relevant. Generally, there are three basic types of relationships of related table: one-to-one, one-to-many, and many-to-many. In our case, these databases belong to one-to-many relationships.

2.1.3.3 Development of Database Management System (DBMS)

By using relational database construction software (e.g. Oracle, Microsoft SQL Server) or even the personal database systems (e.g. MS Access, Fox), data stored in a database can be effectively organized and managed. This kind of data storage and retrieval system is called Database Management System (DBMS). In this work, MS Access DBMSs were used to define, create, maintain and provide controlled access to our databases and repository. All entry data from structured tables described in previous section are brought together for user display and output using SQL queries.

2.2 High throughput QSAR models for virtual screening of drug hits

The process of developing a QSAR model starts with the collection of high quality activity data and the elimination of low quality ones that are likely to affect the accuracy of the model. The next step is the selection of representative compounds into a training set and validation sets to calibrate and evaluate the QSAR model respectively. Molecular descriptors are then computed for representing the physicochemical and structural properties of the compounds studied, and those that are redundant or contain little information are removed prior to the modeling process. Regression methods, in this study the Support Vector Regression (SVR) method, are then used to develop a model that relates the investigated activities of the compounds to their physicochemical and structural properties.

2.2.1 Data preparation

Generally speaking, the performance of QSAR models largely depends on the chemical data quality and diversity of chemical data coverage in the training sets, thus the employment of a systematical chemical record preparation protocol would be helpful in the pre-processing of the chemical dataset.[157] This data preparation process includes high quality data collection, chemical structure (and when possible, associated biological data) curation, and adequate representation of active and inactive chemicals in training datasets.

2.2.1.1 Data source

Data accessibility is critical for the success of a drug discovery and development. Huge amounts of small molecules and their related information have been accumulated in scientific literature and databases. Some important small molecule databases are given in **Table 2.1**.

In this work, datasets including chemical structures and interested biological activities e.g. IC_{50}, EC_{50}, Ki and etc. are mainly collected from the journals (Bioorganic & Medicinal Chemistry Letters, Bioorganic & Medicinal Chemistry, European Journal of Medicinal Chemistry, European Journal of Organic Chemistry and Journal of Medicinal Chemistry,etc) and databases (ChEMBL[158], BindingDB[159], MDDR, PubChem[160] and ZINC[161], etc.).

Table 2.1 Some small molecule databases available online

Database Name	**URL**
BindingDB	http://www.bindingdb.org/bind/index.jsp
MDDR	http://accelrys.com/products/databases/bioactivity/mddr.html
PubChem	http://pubchem.ncbi.nlm.nih.gov/
ZINC	http://zinc.docking.org/
ChEMBL	http://www.ebi.ac.uk/chembl/
DrugBank	http://www.drugbank.ca/
eMolecules	http://www.emolecules.com/
WOMBAT	http://www.sunsetmolecular.com

2.2.1.2 Chemical data curation

Any error in the structure may cause inability to calculate molecular descriptors for erroneous chemical records or resulted in erroneous molecular descriptors. QSAR models developed with these incomplete or inaccurate molecular descriptors may be applicable to only a fraction of available data or even make the models inaccurate.[157] The simple, but important, steps for cleaning chemical records in a dataset include the removal of a fraction of the chemical records that cannot be appropriately handled by conventional cheminformatics techniques, e.g. inorganic and organometallic compounds, counterions, salts and mixtures; structure validation; ring aromatization; normalization of specific chemotypes; curation of tautomeric forms; and the deletion of duplicates and outliers[157]. In this study, the 2D structure of each of the compounds was generated by using ChemDraw or downloaded from other database like PubChem, BindingDB,[159] ChEMBL and etc. and was subsequently converted into 3D structure by using CORINA.[162] All the generated geometries had been fully optimized without symmetry restrictions. The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent was properly generated. All salts and elements, such as sodium or calcium, were removed prior to descriptor calculation.

The development of reliable pharmacological property QSAR models also depends on the availability of high quality pharmacological property descriptor data with low experimental errors.[163] Ideally, these pharmacological properties descriptors should be measured by a single protocol so that different compounds can be reliably compared with each other. However, some pharmacological property descriptors have been measured only for a limited number of compounds and these data are rarely determined by the same protocol. Thus data selection has been primarily based on comparison of data of compounds commonly studied by different protocols, and incorporation of additional experimental information. For this work, several methods are adopted to ensure that inter-laboratory variations in experimental protocols do not significantly affect the quality of the training sets. The sources for the pharmacological property descriptor data for each compound were investigated to remove the chemical records with extreme property descriptors and to ensure that there were no wide variations in experimental protocols from those of the majority of the compounds in the training set. Compounds that were investigated in more than one source are used to estimate the quality of each source.

2.2.1.3 Generation of putative inactive compounds

Active datasets could be generated from available active datasets of sufficiently high number of known actives and varying degrees of structural diversity. On the other hand, putative inactive datasets could be generated by extracting representative compounds from all compound families that contain no known active compound.[164] Compound families can be generated by clustering distinct compounds of chemical databases into groups of similar structural and physicochemical properties.

Apart from the use of known inactive compounds and active compounds of other biological target classes as putative inactive compounds,[165-172] a new approach extensively used for generating inactive proteins in SVM classification of various functional classes of proteins[173-175] has recently been applied for generating putative inactive compounds.[176] An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the "inactive" chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. In applying this approach to proteins, all known proteins are clustered into ~8,933 protein domain families based on the clustering of their amino acid sequences,[177] and a set of putative inactive proteins can be tentatively extracted from a few representative proteins in those families without a single known active protein. By using this method, a reasonably good SVM classification model can be derived from these putative inactive samples, which has been confirmed by a number of studies of proteins.[173-175, 178]

In a similar manner, known compounds can be grouped into compound families by clustering them in the chemical space (PubChem database) defined by their molecular descriptors.[179, 180] As SVR QSAR predict compound activities based on their molecular descriptors, it makes sense to cluster as well as to represent compounds in terms of molecular descriptors. By using a K-means method[179, 180] and molecular descriptors computed from our own software,[181] we generated 8,423 compound families from the available compounds in the PubChem database that we were able to compute the molecular descriptors, which is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms,[182] and the 2,851 clusters for 171,045 natural products.[183]

The collected active compounds could be distributed in hundreds of the 8,423 families. The rest of the families could be taken as inactive datasets candidates and the inactive training dataset corresponding to each sparse or biased active training dataset was generated by random selection of 5~6 representative compounds from each of these "inactive" families and those active families with none of their members in the active training set. The remaining compounds of the "inactive" families in PubChem can be used as putative inactive testing sets. Because of the extensive effort in searching the known compound libraries for identifying active compounds in target classes, the number of undiscovered "active" families in PubChem database is expected to be relatively small, most likely no more than several hundred families. The ratio of the undiscovered "active" families (hundreds on less) and the families that contain no known active compound (7,000~8,000 based on current version of PubChem) for many target classes is expected to be <15%. Therefore, putative inactive compounds can be generated by extracting a few representative compounds of those families that contain no known active compound, with a maximum possible "wrong" family representation rate of <15% even when all of the undiscovered active compounds are misplaced into the inactive class.

2.2.2 Molecular descriptors

Molecular descriptors are generated by a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment. They quantitatively represent structural and physicochemical features of molecules which enables the statistical analysis of chemical compounds.

2.2.2.1 Definition and calculation of molecular descriptors

Molecular descriptors have been extensively used in deriving structure-activity relationships,[184, 185] quantitative structure activity relationships,[186, 187] and machine learning prediction models for pharmaceutical agents.[188-191] A descriptor is "the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a compound into a useful number or the result of some standardized experiment". A number of programs e.g. DRAGON,[192] Molconn-Z,[193] MODEL,[194] Chemistry Development Kit (CDK),[195, 196] JOELib[177] and Xue descriptor set[197] are available to calculate chemical descriptors. These methods can be used for deriving >3,000 molecular descriptors including constitutional descriptors, topological descriptors, RDF descriptors,[198] molecular walk counts,[199] 3D-MoRSE descriptors,[200] BCUT descriptors,[201] WHIM descriptors,[202] Galvez topological charge indices and charge descriptors,[203] GETAWAY descriptors,[204] 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices,[205] Randic molecular profiles,[206] electrotopological state descriptors,[207] linear solvation energy relationship descriptors,[208] and other empirical and molecular properties. Not all of the available descriptors are needed for representing features of a particular class of compounds. Moreover, without properly selecting the appropriate set of descriptors, the performance of a developed ML VS tool may be affected to some degrees because of the noise arising from the high redundancy and overlapping of the available descriptors. In this work, the Xue descriptor set and 98 1D and 2D descriptors were used. These 98 descriptors were selected from the descriptors derived from MODEL program by discarding those that were redundant and unrelated to the problem studied here. The Xue descriptor set and these 98 descriptors are listed in **Table 2.2** and **Table 2.3.**

Table 2.2 Xue descriptor set generated by MODEL program

Descriptor Class	Number of descriptor in class	Descriptors
Simple molecular properties	18	Molecular weight, Number of rings, rotatable bonds, H-bond donors, and H-bond acceptors, Element counts
Molecular connectivity and shape	28	Molecular connectivity indices, Valence molecular connectivity indices, Molecular shape Kappa indices, Kappa alpha indices, flexibility index
Electro-topological state	97	Electrotopological state indices, and Atom type electrotopological state indices, Weiner Index, Centric Index, Altenburg Index, Balaban Index, Harary Number, Schultz Index, PetitJohn R2 Index, PetitJohn D2 Index, Mean Distance Index, PetitJohn I2 Index, Information Weiner, Balaban RMSD Index, Graph Distance Index
Quantum chemical properties	31	Polarizability index, Hydrogen bond acceptor basicity (covalent HBAB), Hydrogen bond donor acidity (covalent HBDA), Molecular dipole moment, Absolute hardness, Softness, Ionization potential, Electron affinity, Chemical potential, Electronegativity index, Electrophilicity index, Most positive charge on H, C, N, O atoms, Most negative charge on H, C, N, O atoms, Most positive and negative

		charge in a molecule, Sum of squares of charges on H,C,N,O and all atoms, Mean of positive charges, Mean of negative charges, Mean absolute charge, Relative positive charge, Relative negative charge
Geometrical properties	25	Length vectors (longest distance, longest third atom, 4th atom), Molecular van der Waals volume, Solvent accessible surface area, Molecular surface area, van der Waals surface area, Polar molecular surface area, Sum of solvent accessible surface areas of positively charged atoms, Sum of solvent accessible surface areas of negatively charged atoms, Sum of charge weighted solvent accessible surface areas of positively charged atoms, Sum of charge weighted solvent accessible surface areas of negatively charged atoms, Sum of van der Waals surface areas of positively charged atoms, Sum of van der Waals surface areas of negatively charged atoms, Sum of charge weighted van der Waals surface areas of positively charged atoms, Sum of charge weighted van der Waals surface areas of negatively charged atoms, Molecular rugosity, Molecular globularity, Hydrophilic region, Hydrophobic region, Capacity factor, Hydrophilic-Hydrophobic balance, Hydrophilic Intery Moment, Hydrophobic Intery Moment, Amphiphilic Moment

Table 2.3 98 molecular descriptors used in this work

Descriptor Class	No of Descriptors in Class	Descriptors
Simple molecular properties	18	Number of C,N,O,P,S, Number of total atoms, Number of rings, Number of bonds, Number of non-H bonds, Molecular weight,, Number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, Number of 5-member aromatic rings, Number of 6-member aromatic rings, Number of N heterocyclic rings, Number of O heterocyclic rings, Number of S heterocyclic rings.
Chemical properties	3	Sanderson electronegativity, Molecular polarizability, ALogp
Molecular Connectivity and shape	35	Schultz molecular topological index, Gutman molecular topological index, Wiener index, Harary index, Gravitational topological index, Molecular path count of length 1-6, Total path count, Balaban Index J, 0-2th valence connectivity index, 0-2th order delta chi index, Pogliani index, 0-2th Solvation connectivity index, 1-3th order Kier shape index, 1-3th order Kappa alpha shape index, Kier Molecular Flexibility Index, Topological radius, Graph-theoretical shape coefficient, Eccentricity, Centralization, Logp from connectivity.
Electro-topological state	42	Sum of Estate of atom type sCH3, dCH2, ssCH2, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH3, sNH2, ssNH2, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH; Sum of Estate of all heavy atoms, all C atoms, all hetero atoms, Sum of Estate of H-bond acceptors, Sum of H Estate of atom type HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsats, HCsatu, Havin, Sum of H Estate of H-bond donors

2.2.2.2 Scaling of molecular descriptors

Chemical descriptors are normally scaled before they can be employed for machine learning. Scaling of chemical descriptors ensures that each descriptor has an unbiased contribution in creating the prediction models.[209] Scaling can be done by number of ways e.g. auto-scaling, range scaling, Pareto scaling,[210] and feature weighting.[209] In this work, range scaling is used to scale the chemical descriptor data. Range scaling is done by dividing the difference between the descriptor value and the minimum value of that descriptor with the in range of that descriptor:

$$d_{ij}^{scaled} = \frac{d_{ij} - d_{j,min}}{d_{j,max} - d_{j,min}} \qquad (1)$$

Where d_{ij}^{scaled}, d_{ij}, $d_{j,\max}$ and $d_{j,\min}$ are the scale descriptor value of compound *i*, absolute descriptor value of compound *i* , maximum and minimum values of descriptor *j* respectively. The scaled descriptor value falls in the range of 0 and 1.

2.2.3 Support Vector Regression (SVR) method

Given the compounds with their activity data and molecular descriptors, a regression model for QSAR can be constructed using SVR to estimate the targeted values. Following is a description explaining how SVR works.

Suppose we are given training data $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset \chi \times \Re$, where χ denotes the space of the input patterns (molecular descriptors derived from structures of compounds as in this study). In ε-SVR,[211] our goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible. In other words, SVR constructs a “tube” with the radius of ε to involve as many training points in it. In linear cases, the function *f* could be written as

$$f(x) = \langle \omega, x \rangle + b \; with \; \omega \in \chi, b \in \Re \tag{2}$$

Where $\langle -, - \rangle$ denotes the dot product in χ. Flatness in the case of (2) means that one seeks a small ω. One way to ensure this is to minimize the norm, i.e. $\|\omega\|^2 = \langle \omega, \omega \rangle$. We can write this problem as a convex optimization problem:

$$\text{Minimize } \frac{1}{2}\|\omega\|^2 \tag{3}$$

$$\text{Subject to } \begin{cases} y_i - \langle \omega, x_i \rangle - b \le \varepsilon \\ \langle \omega, x_i \rangle + b - y_i \le \varepsilon \end{cases}$$

The tacit assumption in (3) was that such a function *f* actually exists that approximates all pairs (x_i, y_i) with ε precision, or in other words, that the convex optimization problem is feasible. Sometimes, however, this may not be the case, or we also may want to allow for some errors. Analogously to the "soft margin" loss function, one can introduce slack variables ξ_i, ξ_i^* to cope with otherwise infeasible constraints of the optimization problem (3). Hence we arrive at the formulation as following:

$$\text{Minimize } \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \tag{4}$$

$$\text{Subject to } \begin{cases} y_i - \langle \omega, x_i \rangle - b \le \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

The constant $C > 0$ determines the trade-off between the flatness of *f* and the amount up to which deviations larger than ε are tolerated. The formulation above corresponds to dealing with a so called ε–insensitive loss function $|\xi|_\varepsilon$ described by

$$|\xi|_{\varepsilon} := \begin{cases} 0 & if\ |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{5}$$

Figure 2.5 depicts the situation graphically. Only the points outside the shaded region contribute to the cost insofar, as the deviations are penalized in a linear fashion. It turns out that the optimization problem (4) can be solved more easily in its dual formulation. Moreover, the dual formulation provides the key for extending SVR to non-linear functions. Hence a standard dualization method utilizing Langrange multipliers will be used.

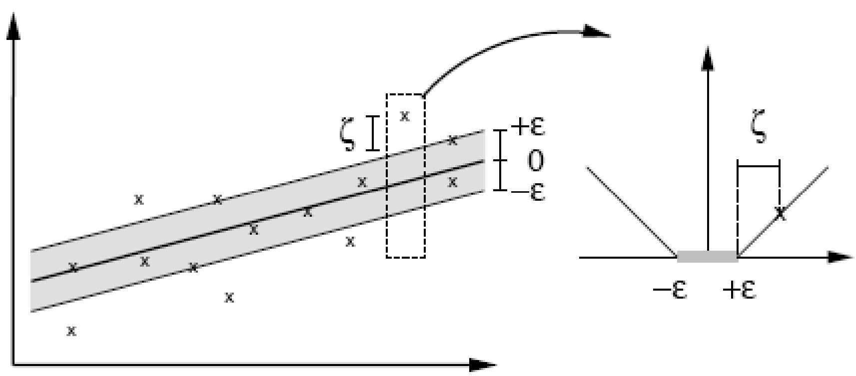


Figure 2.5 The soft margin loss setting corresponds for a linear Support Vector Regression

The key idea is to construct a Lagrange function from both the objective function and the corresponding constraints, by introducing a dual set of variables. It can be shown that this function has a saddle point with respect to the primal and dual variables at the optimal solution. Hence we proceed as follows:

$$L := \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{l}\alpha_i\left(\varepsilon + \xi_i - y_i + \langle \omega, x_i \rangle + b\right)$$

$$- \sum_{i=1}^{l}\alpha_i^*\left(\varepsilon + \xi_i^* + y_i - \langle \omega, x_i \rangle - b\right) - \sum_{i=1}^{l}\left(\eta_i \xi_i + \eta_i^* \xi_i^*\right) \tag{6}$$

It is understood that the dual variables in (6) have to satisfy positivity constraints, i.e. $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$. It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables $(\omega, b, \xi_i, \xi_i^*)$ have to vanish for optimality.

$$\partial_b L = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 \tag{7}$$

$$\partial_\omega L = \omega - \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i = 0 \tag{8}$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} \tag{9}$$

Substituting (7), (8), and (9) into (6) yields the dual optimization problem.

$$\text{Maximize} \begin{cases} -\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*) \end{cases} \tag{10}$$

$$\text{Subject to} \begin{cases} \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

In deriving (10), the dual variables η_i, η_i^* have already been eliminated through condition (9), as these variables did not appear in the dual objective function anymore but only were present in the dual feasibility conditions. Thus (8) can be written as follows:

$$\omega = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i$$

And therefore

$$f(x)=\sum_{i=1}^{l}\left(\alpha_i-\alpha_i^*\right)\langle x_i,x\rangle+b \qquad (11)$$

This is the so-called Support Vector expansion, i.e. ω can be completely described as a linear combination of the training patterns x_i. In a sense, the complexity of a function's representation by support vectors is independent of the dimensionality of the input space χ, and depends only on the number of support vectors. Moreover, the complete algorithm can be described in terms of dot products between the data.

Meanwhile, b can be computed by exploiting the so called Karush-Kuhn-Tucker (KKT) conditions. These state that at the optimal solution the product between dual variables and constraints has to vanish. In the Support Vector case this means

$$\begin{aligned}\alpha_i\left(\varepsilon+\xi_i-y_i+\langle\omega,x_i\rangle+b\right)&=0\\ \alpha_i^*\left(\varepsilon+\xi_i^*+y_i-\langle\omega,x_i\rangle-b\right)&=0\end{aligned} \qquad (12)$$

And

$$\begin{aligned}\left(C-\alpha_i\right)\xi_i&=0\\ \left(C-\alpha_i^*\right)\xi_i^*&=0\end{aligned} \qquad (13)$$

Hence *b* can be computed as follows:

$$\begin{aligned}b&=y_i-\langle\omega,x_i\rangle-\varepsilon \quad for \quad \alpha_i\in(0,C)\\ b&=y_i-\langle\omega,x_i\rangle+\varepsilon \quad for \quad \alpha_i^*\in(0,C)\end{aligned} \qquad (14)$$

After the determination of ω and *b*, the targeted values y_i can be estimated from a given vector χ.

In non-linear regression cases, which frequently occur in QSAR model construction involving diverse structures, SVR maps the input vectors into a higher dimensional feature space by using a kernel function $K(x_i, y_i)$. The mapping mechanism of SVR is constant with the cases in SVM that have been extensively described in previous literature.[212, 213] Thus the details would be skipped here. The kernel function used in this study is the RBF kernel, which has been extensively used and consistently shown better performance than other kernel functions.[214-216]

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2} \quad (15)$$

Linear SVR can then applied to this feature space based on the following decision function:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (16)$$

2.2.4 Tanimoto similarity searching method

Compounds similar to at least one compound in a training dataset can be identified by using the Tanimoto coefficient *sim(i,j)*[217]

$$sim(i, j) = \frac{\sum_{d=1}^{l} x_{di} x_{dj}}{\sum_{d=1}^{l} (x_{di})^2 + \sum_{d=1}^{l} (x_{dj})^2 - \sum_{d=1}^{l} x_{di} x_{dj}} \quad (17)$$

where *l* is the number of molecular descriptors. A compound *i* is considered to be similar to a known active j in the active dataset if the corresponding *sim(i,j)* value is greater than a cut-off value. In this work, the similarity search was conducted for MDDR compounds. Therefore, in computing *sim(i,j)*, the molecular descriptor vectors $\mathbf{x}_i$'s were scaled with respect to all of the

MDDR compounds. The cut-off values for similarity compounds are typically in the range of 0.8 to 0.9.[218, 219] A stricter cut-off value of 0.9 was used in this work.

2.2.5 Model validation and virtual screening performance evaluation

Derived from application of statistical tools correlating biological activity of chemicals with descriptors representative of molecular structure and/or properties, QSAR models can then be adapted for lead optimization and modification, and virtual screening large chemical database for novel drug hits. Obtaining a good quality QSAR model depends on many factors, such as the quality of biological data as described in **Section 2.1.1**, the choice of descriptors and statistical methods. Any QSAR modeling should ultimately lead to statistically robust models capable of making accurate and reliable predictions of biological activities of new compounds. In this work, the QSAR models are evaluated by adopting three strategies: internal 5-fold cross-validation, external test validation and evaluation on performance for large chemical database virtual screening.

2.2.5.1 Internal 5-fold cross-validation

In 5-fold cross-validation, the curated collection of compounds is randomly partitioned into 5 subsets. Of the 5 subsets, each single subset is retained as the validation data for testing the model, and the remaining 4 subsets are used as training data. The cross-validation process is then repeated for 5 times. The squared cross-validation correlation coefficient Q^2 is employed for evaluating the internal predictivity of QSAR models.

$$Q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where y_i is the experimentally observed activity for each compound, $\hat{y}_i$ is the *in-silico* determined activity from cross-validation, and $\bar{y}$ is the averaged observed activity of all compounds included in all the 5 folds.

2.2.5.2 External/independent test validation

In a very important review paper entitled "*Beware of q^2!*"[220] Golbraikh and Tropsha demonstrated that the high accuracy of the traning set model characterized with leave-one-out (LOO) or leave-some-out cross validated q^2 is not indicative of the high external predictive power of the model. Thus QSAR models exclusively relying on training set modeling without any external validation are bad at generalization and considered to be unreliable.

In developing our SVR QSAR model, a hard margin C=1,000 was used and the predictivity of the model on external test set is evaluated by the Correlation Coefficient (R) and Mean Squared Error (MSE).

$$R = \left(\sum y_i \hat{y}_i - \frac{\sum y_i \cdot \sum \hat{y}_i}{n} \right) \Bigg/ \sqrt{\left[\sum y_i^{\ 2} - \frac{\left(\sum y_i \right)^2}{n} \right] \left[\sum \hat{y}_i^{\ 2} - \frac{\left(\sum \hat{y}_i \right)^2}{n} \right]}$$

$$MSE = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n}$$

where y_i is the actual activity measured by experiments in testing datasets, $\hat{y}_i$ denotes the estimated value and n is the total number of compounds in testing dataset.

2.2.5.3 Performance evaluation on large chemical database virtual screening

The typical measurements of a model performance in screening large libraries include[221] yield (percentage of known positives predicted as virtual hits), hit-rate (percentage of virtual hits that are known positives), false hit-rate (percentage of virtual hits that are known negatives) and enrichment factor EF (magnitude of hit-rate improvement over random selection):

$$yield = sensitivity\ (SE) = \frac{TP}{TP + FN}$$

$$Hit\ rate = \frac{TP}{(TP+FP)}$$

$$False\ hit\ rate = \frac{FP}{TP + FP}$$

$$Enrichment\ factor\ (EF) = \frac{hit\ rate}{(TP + FN)/(TP + FN + TN + FP)}$$

2.2.6 Overfitting problem and its detection

Overfitting is a major concern in machine learning regression methods. It happens when a model that agrees well with the observed data but has no predictive ability, which means it does not have any value to unseen or future data. There are two main types of overfitting situations: (1) a model more flexible than it needs to be and (2) a model including irrelevant descriptors.[222] An over-fitted classification system tends to obtain much higher prediction accuracies in the cross-validation sets than in the independent validation sets. Hence frequently used method for checking whether a model is overfitted is to compare the prediction accuracies in the cross-validation procedure with those found in testing independent validation sets.[222]

CHAPTER 3 Development of Pathway Cross-talk Database Facilitating Multi-target Selection

Cross-talk between pathways plays important regulatory roles in biological processes, disease processes, and therapeutic responses. Knowledge of these cross-talks is highly useful for facilitating systems level analysis of diseases, biological processes and the mechanisms of multi-targeting drugs and drug combinations. However, to our best knowledge, currently no such database exists providing this kind of information. Developed in the year of 2008, the Pathway Cross-talk Database (PCD) provides information about experimentally discovered cross-talks between pathways and their relevance to diseases and biological processes, mechanism of multi-target drugs and drug combinations. In this chapter, the data source, structure and access of PCD are introduced in details. The usefulness of PCD in facilitating systems level studies of diseases and mechanism of drug combinations and multi-targeting drugs is demonstrated by the analysis of the effect of glutamate on glioma cell invasion, the synergistic actions of tamoxifen-herceptin drug combination, and multi-targeting cross-talked signaling pathways, e.g. EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src pathways, as the prospective direction for treating non-small cell lung cancer (NSCLC).

3.1 Introduction

Biological pathways are part of biological systems that play context-dependent and specific metabolic and signal transduction tasks, and cross-talks between these pathways facilitate the regulation and coordination of biomolecular events in biological, disease, and therapeutic processes in responses to internal changes, external stimuli, and actions of therapeutic agents.[223-225] Individual pathways alone cannot fully represent signaling networks of the cell and methods for collective analysis of the dynamics of multiple network elements have been developed.[226, 227] None-the-less, individual pathway concept and the relevant models are useful building blocks for more comprehensive understanding of network collective actions, and knowledge of pathway cross-talks further facilitates and extends the use of the pathway concept for studying biological[228-230] and disease[231-237] processes, for discovering multi-targeting drugs and drug combinations,[1, 18, 123, 238, 239] and for simulating and theoretically investigating the biological events.[240, 241]

A number of pathway databases have been developed to provide comprehensive information about the molecular interactions and networks of a variety of metabolic, transport, and signaling pathways.[242-246] Experimental studies have shown the existence of cross-talk between many different pathways. Our search of literature identified 137 experimentally discovered pathway cross-talks among 89 pathways or pathway components with sufficient information about the molecular interactions or regulations that mediate these cross-talks. The relevant information has not been specifically provided in the existing pathway databases. Databases of protein functional association networks such as STRING[247] and Reactome[248] are useful resources for assessing interactions that may mediate some of the reported cross-talks, However, these databases are not specifically designed for convenient access of cross-talk interactions, and some of the interactions, particularly those via regulation of protein levels, have not been included in these databases.

A public resource for providing the relevant information about these and other pathway cross-talks is helpful in complementing and expanding the application scope of the existing pathway and protein association databases. A new database, Pathway Cross-talk Database (PCD), was introduced as a public resource of experimentally discovered pathway cross-talks. PCD provides detailed description about cross-talking partners, their mediators in terms of molecular interactions or regulations, cross-talk effects, related diseases or biological processes, and relevant references. Pathway maps and graphical representation of the cross-talks are provided in PCD. Cross-links to other databases, including NCBI,[249] KEGG,[242] SwissProt,[250] BioCarta,[24] Ambion,[251] and Cell Signaling Technology,[252] are provided to further facilitate the access of network maps and other information.

3.2 Database information source, structure and access

PCD has a web interface at http://bidd.nus.edu.sg/group/PCD/PCD.asp, which is shown in **Figure 3.1**. The entries of this database were generated from a comprehensive search of published literature via PubMed by using a similar search and inspection procedure as we have used for developing databases of functional proteins and effects.[253-257] We used the keyword “crosstalk” combined with either “pathway” or “network” or “protein” to identify the literature that describe experimentally discovered cross-talk between two different pathways. A total of 650, 170, and 1,022 abstracts were obtained by the keyword search, which were reduced to 447 entries after removing redundant and irrelevant entries. Irrelevant entries are those describing inter-cellular, inter-tissue, or intra-pathway cross-talks. These 447 literature were further inspected manually to select 137 entries with sufficiently detailed information about the molecular interactions or regulations mediating the cross-talk. Members of each pathway were retrieved from Ambion[251] and Biocarta[258] databases, and the corresponding protein and gene IDs were retrieved from SwissProt database.[250]

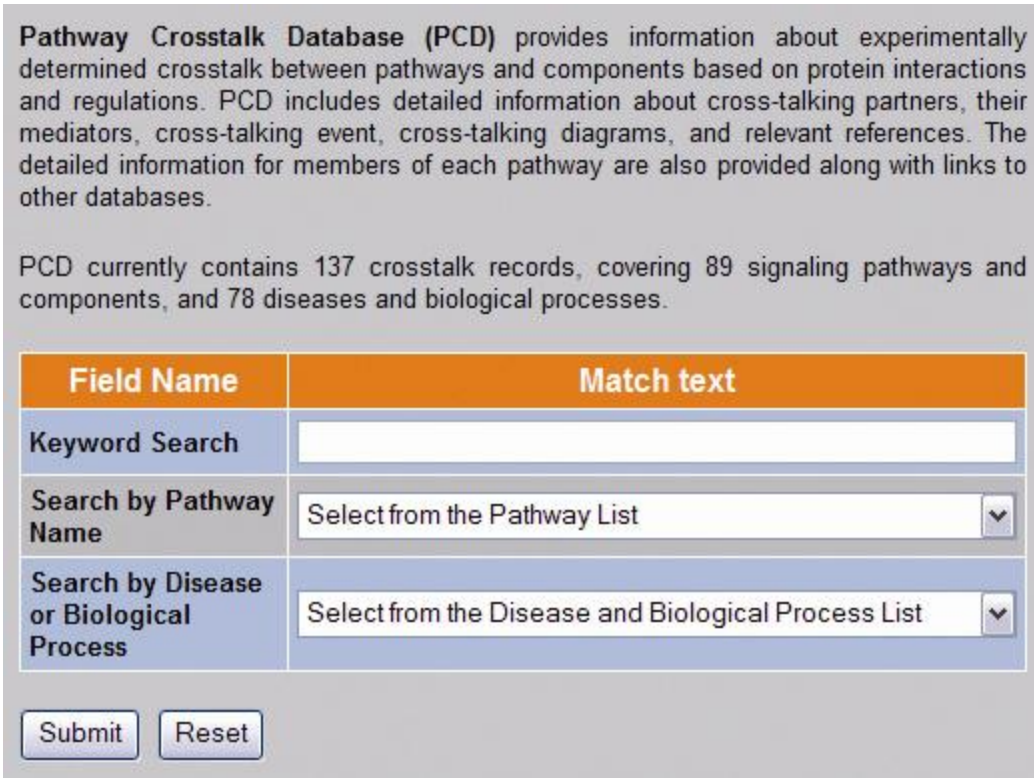
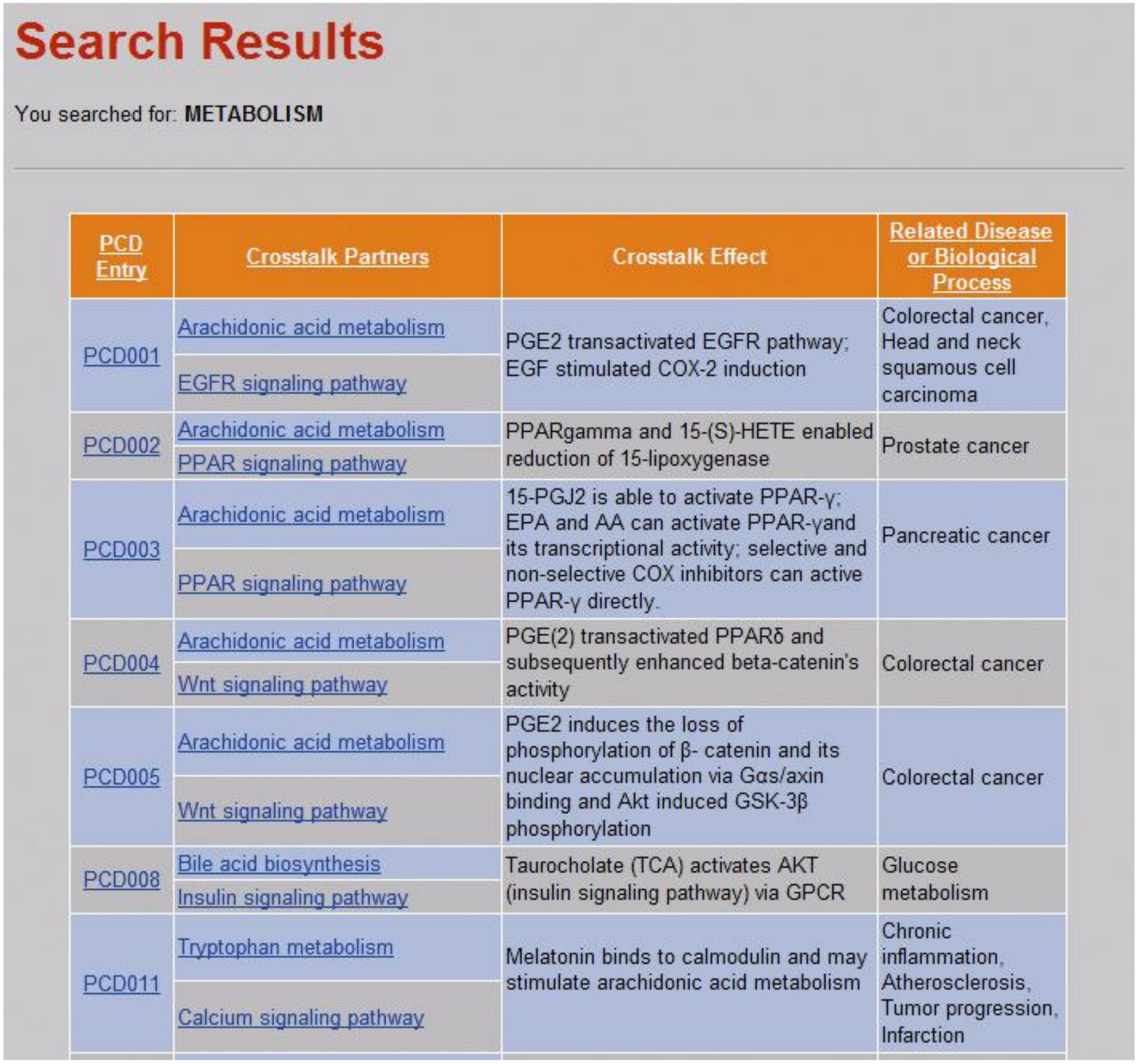


Figure 3.1 Web-page of PCD

PCD is browseable and searchable via the names and list of cross-talk pathways or pathway components and via the names and list of disease or biological processes provided in the PCD webpage. Download and keyword search are also supported via download link and keyword search window in the webpage.

The cross-talk entries are browseable and searchable via the names and list of cross-talk pathways or pathway components and via the names and list of disease or biological processes provided in the PCD webpage (**Figure 3.1**). Download and keyword search are also supported via download link and keyword search window in the webpage. The result of a typical search is illustrated in **Figure 3.2**, in which all cross-talks that satisfy the search criteria are listed. This list includes the names of the cross-talk pathways and links to each cross-talk entry. These entries can be ordered by name of cross-talk partner (pathway), disease name, and PCD entry ID. The detailed information related to a distinct entry can be obtained by clicking the PCD entry ID of a selected cross-talk. The page of a cross-talk entry, as shown in **Figure 3.3**, provides detailed description about the names of cross-talk pathways or pathway components, cross-talk mediator in terms of molecular interactions or regulations, cross-talk effect, related diseases or biological processes, literature descriptions, related references, and cross-talk map (an example can be seen in **Figure 3.4**). Further information about the maps and protein members of the cross-talk pathways can be obtained by clicking the name of the respective pathway. As shown in **Figure 3.5**, the corresponding pathway information page provides the pathway map and links to one or more of the pathway databases KEGG,[242] BioCarta,[258] Ambion,[251] and Cell Signaling Technology[252] in which further information of the pathway is available. Enzyme or protein information such as enzyme name and catalyzed reaction or protein name, gene name, SwissProt accession number, and amino acid sequence for each member of the pathway or pathway component can also be retrieved by clicking the corresponding component block in the map.



Search Results

You searched for: **METABOLISM**

PCD Entry	Crosstalk Partners	Crosstalk Effect	Related Disease or Biological Process
PCD001	Arachidonic acid metabolism EGFR signaling pathway	PGE2 transactivated EGFR pathway; EGF stimulated COX-2 induction	Colorectal cancer, Head and neck squamous cell carcinoma
PCD002	Arachidonic acid metabolism PPAR signaling pathway	PPARgamma and 15-(S)-HETE enabled reduction of 15-lipoxygenase	Prostate cancer
PCD003	Arachidonic acid metabolism PPAR signaling pathway	15-PGJ2 is able to activate PPAR-γ; EPA and AA can activate PPAR-γand its transcriptional activity; selective and non-selective COX inhibitors can active PPAR-γ directly.	Pancreatic cancer
PCD004	Arachidonic acid metabolism Wnt signaling pathway	PGE(2) transactivated PPARδ and subsequently enhanced beta-catenin's activity	Colorectal cancer
PCD005	Arachidonic acid metabolism Wnt signaling pathway	PGE2 induces the loss of phosphorylation of β- catenin and its nuclear accumulation via Gαs/axin binding and Akt induced GSK-3β phosphorylation	Colorectal cancer
PCD008	Bile acid biosynthesis Insulin signaling pathway	Taurocholate (TCA) activates AKT (insulin signaling pathway) via GPCR	Glucose metabolism
PCD011	Tryptophan metabolism Calcium signaling pathway	Melatonin binds to calmodulin and may stimulate arachidonic acid metabolism	Chronic inflammation, Atherosclerosis, Tumor progression, Infarction

Figure 3.2 The interface for a search in PCD

All entries that match the search selection are listed. This list includes the name of cross-talk partners, brief description of cross-talk effects, related diseases or biological processes, and entry access to the detailed cross-talk information.

Detailed Information for PCD002

Field Name	Content
Crosstalk Partner A	Arachidonic acid metabolism
Crosstalk Partner B	PPAR signaling pathway
Crosstalk Mediator	15-Lipoxygenase-2 (15-LOX-2) product 15-(S)-HETE binds to PPARgamma, the complex subsequently binds to 15-LOX-2 promotor to reduce the expression of 15-LOX-2
Crosstalk Effect	PPARgamma and 15-(S)-HETE enabled reduction of 15-lipoxygenase
Literature Descriptions	An inverse relationship exists between the expression of 15-LOX-2 and PPARgamma in normal prostate ecithelial cells (PrECs) compared with their expression in prostate carcinoma cells (PC-3). This inverse expression partly involves the 15-LOX-2 promoter and 15-(S)-HETE, a product of 15-LOX-2 that binds to PPARgamma. We identified an active steroid nuclear receptor half-site present in the 15-LOX-2 promoter fragment F-5 that can interact with PPARgamma. After forced expression of wild-type PPARgamma, 15-(S)-HETE decreased F-5 reporter activity in PrECs whereas forced expression of 15-LOX-2 resulted in 15-(S)-HETE production which enhanced F-5 activity in PC-3. In contrast, the expression of dominant-negative PPARgamma reversed the transcriptional activation of F-5 by enhancing it 202-fold in PrEC or suppressing it in PC-3; the effect in PC-3 was positively increased 150-fold in the presence of 15-(S)-HETE.
Related Disease or Process	Prostate cancer
Reference	1. Subbarayan V, Krieg P, Hsi LC, Kim J, Yang P, Sabichi AL, Llansa N, Mendoza G, Logothetis CJ, Newman RA, Lippman SM, Menter DG. 15-Lipoxygenase-2 gene regulation by its product 15-(S)-hydroxyeicosatetraenoic acid through a negative feedback mechanism that involves peroxisome proliferator-activated receptor gamma. Oncogene. 2006 Sep 28;25(44):6015-25. PubMed
Crosstalk Map	 Click to view big map

Figure 3.3 Cross-talk information page

This page provides information about cross-talk partners, cross-talk mediator in terms of molecular interactions or regulations, cross-talk effect, related diseases or biological processes, the detailed description in literature and references as well as the graphical representation of the cross-talk. Further information about the participating components can be obtained by clicking the name of these components.

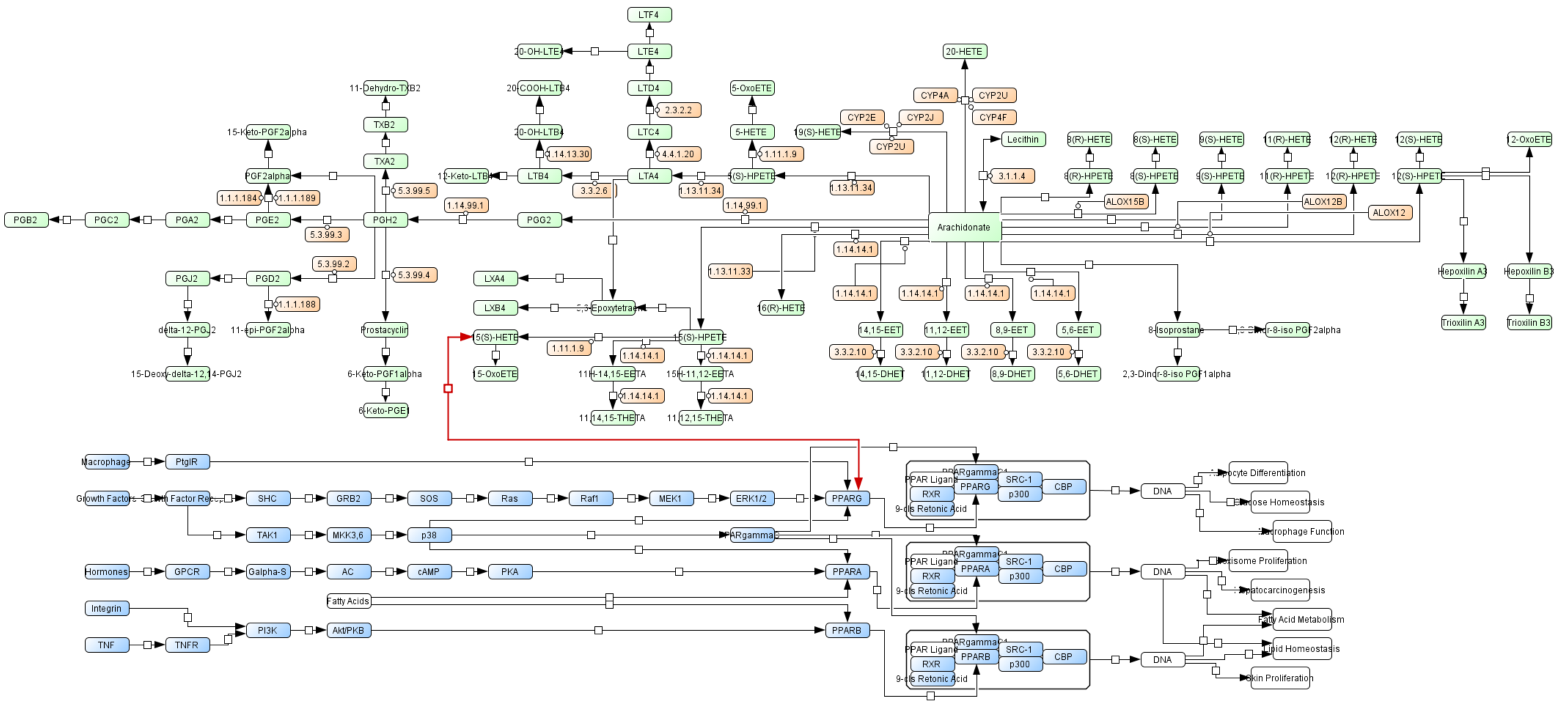


Figure 3.4 An example of graphical representation for pathway cross-talk. Cross-talk between Arachidonic acid metabolism and PPAR signaling pathway

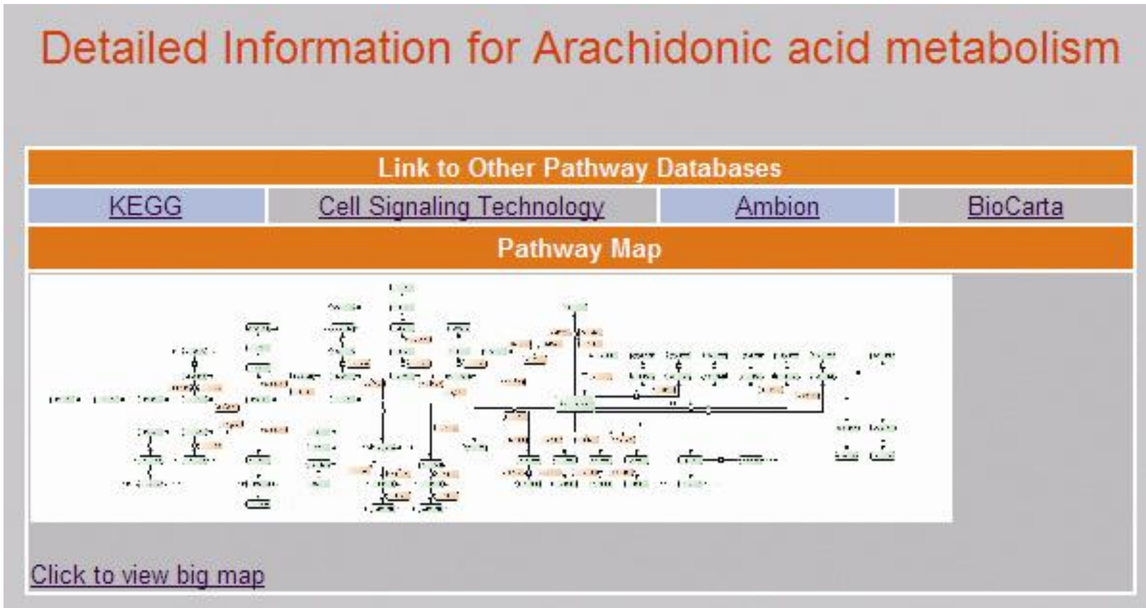


Figure 3.5 Pathway information page

Pathway maps and links to other four common-used pathway databases – KEGG, Cell Signaling Technology, Ambion, and BioCarta – have been provided for each pathway covered by PCD. Detailed information for each member of the pathway or pathway component can also be retrieved by clicking the corresponding component block in the map.

3.3 Potential applications of PCD

3.3.1 Systems level analysis of diseases

One potential application of PCD in facilitating systems level study of diseases can be illustrated by the analysis of recently discovered effects of glutamate signaling in promoting glioma cell invasion.[259] Malignant gliomas have been shown to release glutamate that kills surrounding brain cells, creating room for tumor expansion. This glutamate release occurs primarily via system xC, a Na+-independent cystine-glutamate exchanger. The released glutamate also acts as an essential autocrine/paracrine signal that promotes cell invasion.[259] The mechanism of glutamate promotion of cell invasion can be partly explained by the cross-talk between Ca2+-permeable AMPA receptor pathway and PI3K-AkT Pathway. In Bergmann glia, glutamate binding to Ca2+-

permeable AMPA receptors leads to receptor tyrosine phosphorylation, which subsequently interacts and activates PI3K, activated PI3K then activates AkT leading to the promotion of cell invasion.[260] Therefore, Akt functions as downstream effectors for Ca2+-signaling mediated by AMPA receptor in glioblastoma cells. AkT activation via the cross-talk between glutamate-AMPA receptor pathway and PI3K-Akt pathway may contribute to the high degree of anaplasia and invasive growth of human glioblastoma.[261]

3.3.2 Systems level analysis of synergistic drug combinations

The potential application of PCD can be further illustrated by the analysis of literature-reported synergistic drug combinations. Tamoxifen-Trastuzumab (Herceptin) combination has been found to synergistically inhibit the growth in ER- positive, HER-2/neu overexpressing BT-474 breast tumor cells.[262, 263] Tamoxifen is an estrogen receptor (ER) antagonist[264] and herceptin is an anti-HER-2/neu antibody[265] extensively used for the treatment of breast cancers. The synergistic actions of this drug combination can be partly explained by their collective regulation of the cross-talk between estrogen receptor pathway and HER2 signaling. HER2 is known to activate p42/44 MAPK, which subsequently activates ER and ER coactivator AIB1.[266-268] Moreover, ER directly interacts with HER2 in the membrane to transactivate HER2 and its signaling.[269] Apart from inhibiting HER-2 signaling, the anti-HER-2/neu antibody herceptin stops HER-2/neu induced activation of ER and AIB1. On the other hand, ER antagonist tamoxifen stops ER induced transactivation of HER-2, leading to synergistic actions.

3.3.3 Systems level analysis of multi-targeting drugs and multi-target selection

Another potential application of PCD can be illustrated by the systems level study of multi-target agents to achieve enhanced therapeutic efficacies and reduced drug resistance activities by targeting multiple interacted signaling pathways. One example can be demonstrated by the necessity of the collective inhibition of EGFR and VEGFR signaling pathways in treating non-small cell lung cancer (NSCLC). NSCLC is the most common type of lung cancer that is responsible for the highest number of cancer deaths.[270] Because lung cancer is typically diagnosed at an advanced stage, the prognosis and survival rate for patients are poor and have remained not improved for decades.[271] Targeted inhibition of either EGFR or VEGFR signaling pathways has been clinically validated in advanced NSCLC with a number of approved drugs e.g. bevacizumab (Avastin), erlotinib (Tarceva), cetuximab (Erbitux) and gefitinib (Iressa). However, in some cases, these drugs exhibit moderate efficacies, undesired AEs and resistance profiles.[2] The acquired resistance can be partially attributed to the cross-talks between EGFR and VEGFR signaling pathways in which the VEGF can be up-regulated independent of EGFR signaling thus promoting tumor angiogenesis.[2, 6, 10] The detailed description of one possible mechanism can be found in **Section 1.1.1.**

Besides, the reduced sensitivity to EGFR inhibitors in NSCLC patients may also be linked to acquired alternative routes of proliferative and survival signaling, e.g. PDGFR and FGFR, bypassing EGFR signaling.[272] PDGFR and FGFR are aberrantly expressed in mesenchymal-like NSCLC cells.[272] The autophosphorylation and substrate-phosphorylation of PDGFR has been shown to be significantly increased when EGFR was inhibited.[272] Evidence also showed that FGFR inhibition had an effect on ERK signaling and to a lesser extent on Akt signaling in two mesenchymal-like NSCLC cell lines, H1703 and H226, which were growth inhibited when treated with FGFR inhibitors. These findings suggested that, via PDGFR and FGFR, the autocrine signaling can activate the EGFR downstreamed MEK-ERK and PI3K signaling in an EGFR-independent manner.[272]

Another kinase, Src, has been reported to be increased expressed in 50% of squamous cell carcinomas isolated from patients with NSCLC.[272] In addition, high levels of Src kinase activity have also been reported in NSCLC correlating with enlarged tumor size.[151] Constitutive activation of EGFR is found in a subset of NSCLC tumors that are dependent on EGFR for survival.[148] Besides EGFR, kinase Src also offers a promising target for treating NSCLC since the inhibition of it can lead to the inhibition of multiple signaling pathways including those mediated by EGFR.[148] One possible path is Src activation of EGFR by phosphorylating tyrosine residue Y845 to promote oncogenesis via STAT-5b independent of the ERK2 pathway.[149, 150, 152] And the synergistic effect of EGFR and Src in promoting aggressive phenotype has been evidenced in nude mice that tumors in nude mice inoculated with EGFR/Src overexpressing fibroblasts were significantly larger than those inoculated with fibroblasts overexpressing either EGFR or Src alone.[152]

Therefore, collective blockade of interactive cross-talked signaling pathways or key components of these pathways, e.g. EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src, may offer the treatment for NSCLC with enhanced therapeutic effects. In **Chapter 5,** a novel high throughput SVR QSAR approach is developed and used for searching dual inhibitors of these kinase combinations.

CHAPTER 4 Construction of QSAR Models with Enhanced Ability for Searching Highly Novel Hits

Based on a new chemspace-wide regression strategy, in this chapter, we developed support vector regression (SVR) QSAR models applicable beyond similarity-based applicability domains. In screening large chemical libraries, these QSAR models built from pre-2010 DHFR, ACE and Cox2 inhibitors showed substantial predictive capability for post-2010 and patented inhibitors outside the domains, while performed equally well for inhibitors within the domains as the established QSAR methods.

4.1 Introduction

Apart from drug lead optimization, QSAR models have been developed for searching drug leads, particularly novel ones, from large chemical libraries.[130-137] These models achieve good hit rates and activity assessment by pharmacophoric-shim adjusted molecular docking (PSA-Docking),[130-132] Bayesian-based target-family activity profiling (BTFAP),[133] and machine learning regression (MLR) of known actives[134-137] within applicability domains (ADs) defined by binding-mode constraints,[130] Baysian active-inactive boundaries,[133, 138] and range-based and distance-based similarity to the known actives.[139, 140] In particular, MLR requires no knowledge of target 3D structure or target-family activity profiles,[141] but cannot find highly novel actives outside similarity-based ADs. In this work, we extended an approach in the BTFAP method[133] for constructing new MLR QSAR models via chemspace-wide activity regression aimed at finding highly novel inhibitors without compromising hit rates within similarity-based ADs. Our consensus QSAR models developed by “old” (pre-2010) DHFR, ACE and Cox2 inhibitors performed well in predicting the activities of “new” (post-2010) inhibitors with R^2 values comparable to those of the kNN QSAR,[137] PSA-Docking[130-132] and BTFAP[133] methods, and in identifying inhibitors from large chemical libraries (168,016 MDDR and 13.56 million PubChem compounds) at improved hit rats and enrichment factors. In particular, our method showed some level of capability in the identification and activity assessment of highly novel inhibitors outside similarity-based ADs.

4.2 Materials and methods

4.2.1 Compound collection, training and testing datasets, molecular descriptors

Chemically diverse sets of 760, 803 and 2,467 DHFR, ACE and Cox2 inhibitors ($pIC_{50}>5$) and 200, 127 and 618 non-inhibitors ($pIC_{50}\leq5$) published before 2010 were collected from the ChEMBL database[158] and additional literature search,[164] which were tentatively regarded as “old” inhibitors and non-inhibitors and used for developing QSAR models. From the ChEMBL database, we collected additional sets of 26, 47 and 72 DHFR, ACE and Cox2 inhibitors and 46, 54 and 50 non-inhibitors published since 2010, which were tentatively regarded as “new” inhibitors and used for testing QSAR models. Moreover, the MDDR database contains 167, 532 and 990 DHFR, ACE and Cox2 inhibitors not found in the ChEMBL database, which together with the rest of the 168K MDDR compounds and 13.56 million compounds from the PubChem database[273] were used for testing the ability of QSAR models in the virtual screening (VS) of large chemical libraries. By using the Chembench web-based tool[274] for QSAR modeling and prediction with the parameters adjusted to reproduce the results of the published HDAC inhibitor QSAR screening studies,[137] we found that 0.00%, 12.77% and 18.06% of the post-2010 and 14.97%, 32.89% and 5.15% of the patented DHFR, ACE and Cox2 inhibitors and 8.70%, 62.96% and 14.00% of the post-2010 non-inhibitors are outside the similarity-based ADs derived by the method of the Tropsha group[134, 137] with respect to the pre-2010 inhibitors and non-inhibitors, suggesting that substantial percentage of the "new" inhibitors are highly novel ones outside the typical similarity-based ADs.

While conventional MLR QSAR models are applicable within specific ADs,[275] a method for extending the applicability of MLR QSAR models beyond similarity-based ADs has been outlined by Martin et al in their profile-QSAR modeling of kinase inhibitory activities.[133] In their method,[133] actives of an individual target are divided into specific activity ranges, within each range a Bayesian classification model is developed from the in-range actives and combination of the out-range actives and chemically diverse inactives,[138] a Bayesian QSAR model is subsequently constructed based on the experimental activity values of the compounds (pIC_{50}>4) and a uniformly assigned activity value (pIC_{50}=3) for all inactives with $pIC_{50} \leq 3$ or unknown values. The inclusion of chemically diverse inactives helps refining active-inactive boundaries for enhanced identification of highly novel actives.[138, 164] In this work, we further improved Martin et al's method in three aspects. The first is the significant expansion of the inactive chemspace from one corporate archive (1.5 million) to all Pubchem and MDDR compounds (13.7 million). The second is the chemspace-wide regression of compounds by a single MLR directly based on experimental (for actives) and assigned (for inactives) activity values without dividing actives into specific activity ranges. The third is the assignment of the activity values of putative inactives based on the distance-dependent activity profiles revealed by the regression of the experimental activity values of the known inactives with respect to their closest distances to the known potent actives, instead of assignment of a uniform activity value.

For each target, putative inactives covering Pubchem and MDDR compounds were generated by using our previously-reported method that requires no knowledge of known inactives or actives of other target classes.[164, 276] The 13.56M PubChem and 168K MDDR compounds were clustered into 8,423 compound families by using molecular descriptor Tanimoto similarity coefficients[217]

$$sim(i, j) = \frac{\sum_{d=1}^{l} x_{di} x_{dj}}{\sum_{d=1}^{l} (x_{di})^2 + \sum_{d=1}^{l} (x_{dj})^2 - \sum_{d=1}^{l} x_{di} x_{dj}}$$

where $\{x_{di}, d=1, \ldots, l\}$ are molecular descriptors for the *i*-th compound computed, and the molecular descriptors were computed from the MODEL[194] program. The detailed description of these molecular descriptors can be found in **Section 2.2.2**.

Our collected DHFR, ACE and Cox2 inhibitors are in 76, 188 and 901 families respectively. The numbers of families without a known DHFR, ACE and Cox2 inhibitor are 8,347, 8,235 and 7,522 respectively, which is consistent with the number of 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4M compounds of up to 11 atoms[277] and that of the 2,851 structural clusters for 171,045 natural products.[278] By selecting one representative compound from each family containing no known inhibitor as a putative inactive, we obtained 8,347, 8,235 and 7,522 putative inactives for representing the inactive chemspace of PubChem and MDDR compounds, which were used for training MLR QSAR models. Some new inhibitors are likely distributed in the families whose representative is regarded as a putative inactive, a substantial percentage of these new inhibitors are expected to be identifiable as hits even if their family representatives are regarded as inactives.[164]

To assign activity values of the putative inactives, we derived the distance-dependent pIC_{50} regression profiles of the 30, 68 and 111 known DHFR, ACE and Cox2 non-inhibitors ($2<pIC_{50}<4$) with respect to their closest distances to the 282, 492 and 759 known potent inhibitors ($pIC_{50}>7$) from their experimental activity values and molecular fingerprint Tanimoto similarity coefficients (**Figure 4.1-4.3**), with molecular fingerprints computed by using PaDEL.[279] From these profiles, the pIC_{50} values of the 8,347, 8,235 and 7,522 putative inactives were assigned based on their closest distances to the 282, 492 and 759 known potent inhibitors, which are in the range of 2.87-3.67, 2.48-3.66 and 3.01-3.74 with median values of 3.32, 3.22 and 3.48 that are consistent with Martin et al's assignment of pIC_{50}=3 for inactives.[133]

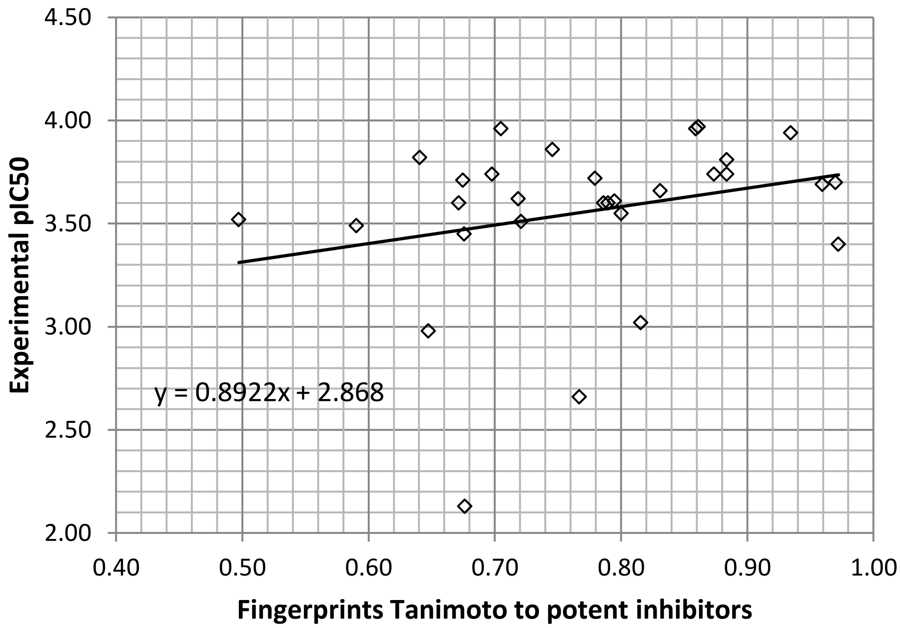


Figure 4.1 The pIC_{50} values of the known DHFR non-inhibitors ($2<pIC_{50}<4$) with respect to their closest distances to the known potent inhibitors

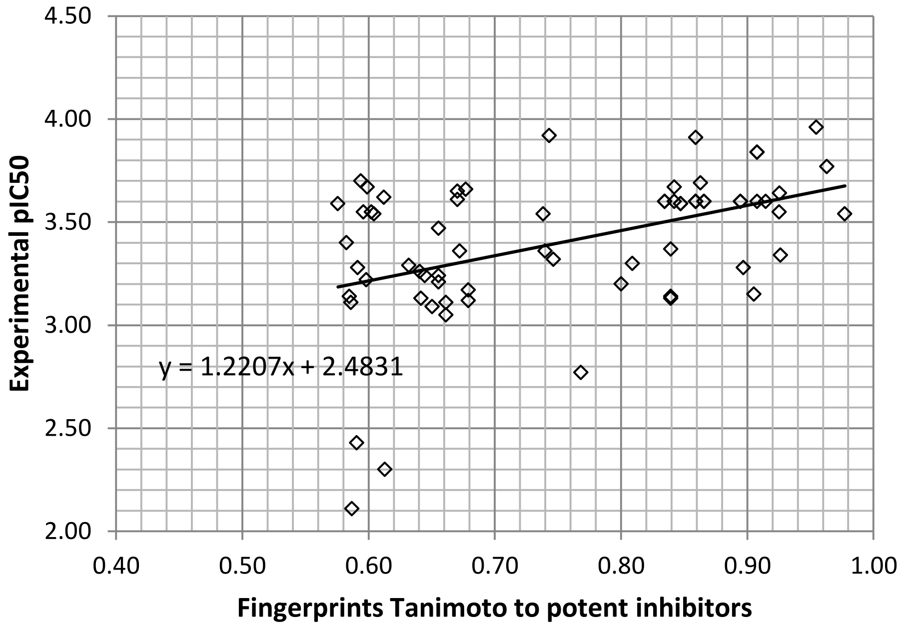


Figure 4.2 The pIC_{50} values of the known ACE non-inhibitors ($2<pIC_{50}<4$) with respect to their closest distances to the known potent inhibitors

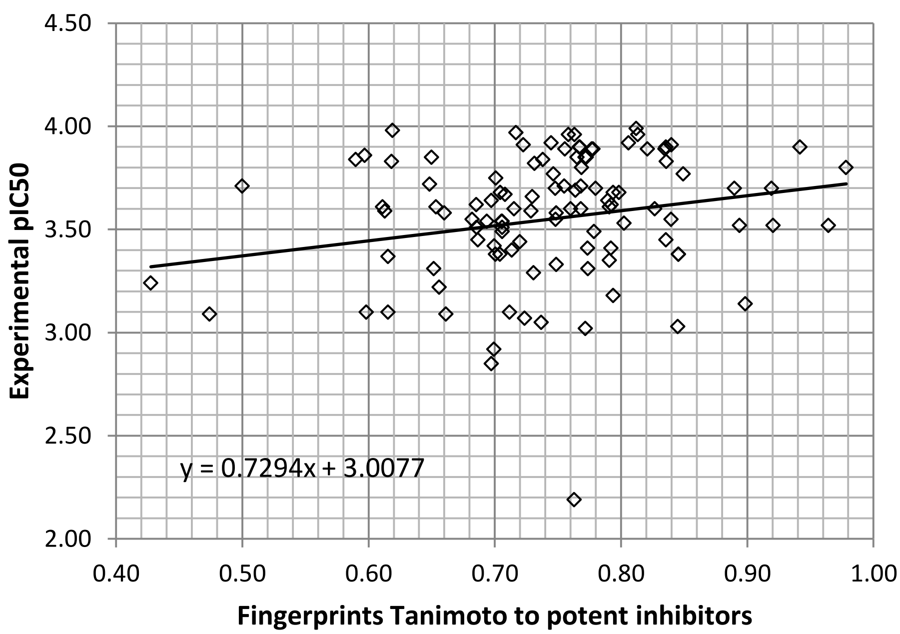


Figure 4.3 The pIC_{50} values of the known Cox2 non-inhibitors ($2<pIC_{50}<4$) with respect to their closest distances to the known potent inhibitors

4.2.2 Computational models

We used a MLR method, support vector regression (SVR), for deriving QSAR models not only because it has consistently shown good performance,[135, 280-285] but also because it is less penalized by sample redundancy and has lower risk for over-fitting.[286, 287] The latter is particularly important for chemspace-wide regression. Given a training dataset $\{(x_1, y_1), \ldots, (x_l, y_l)\}$, where x_i is the input vector composed of molecular descriptors of compound i and y_i is its activity value, the objective of ε-SVR[211] is to find a function $f(x)$ that minimally deviates from the activity values $\{ y_i \}$ of the training compounds (with deviation amplitude less than ε), i.e., it constructs a tube of radius of ε to maximally include training compounds. In linear regression

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b$$

where $\alpha_i \geq 0$ are Lagrange multipliers. In non-linear regression, which frequently occur in developing QSAR from chemically diverse compounds, SVR maps the input vectors into a higher dimensional feature space by using a kernel function $K(x_i, y_i)$. The kernel function used in this study is the RBF kernel,

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2}$$

which has been extensively used and consistently shown better performance than other kernel functions.[214-216] Linear SVR can then be applied to this feature space based on the following decision function.

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

More mathematical details about the SVR method can be found previously in **Section 2.2.3.** The actual QSAR models were developed by using the SVR module of the LIBSVM software[288] with RBF kernel,[214-216] a hard margin C=1,000, and ε =0.19-0.60.

4.3 Results and discussion

4.3.1 Performance of SVR QSAR models in identification of DHFR, ACE and Cox2 inhibitors based on 5-fold cross validation test

The QSAR models for DHFR, ACE and Cox2 inhibitors were trained and tested by using 5-fold cross validation (5-fold CV) method. For each target, training inhibitors and non-inhibitors were randomly divided into 5 groups of approximately equal size, with 4 groups used for training an SVR model and 1 group used for testing it, and the process was repeated for all 5 possible training-testing configurations. The squared correlation coefficient

$$q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

was used for preliminary performance evaluating of the QSAR models, where y_i and $\hat{y}_i$ are the actual and predicted activity value of compound i, and $\bar{y}$ is the average predicted activity value of all compounds over all 5 folds. For each target, the top 15 SVR QSAR models with the best 5-fold CV performance (**Table 4.1-4.3**) were used for constructing a consensus SVR QSAR model for testing their ability in identifying “new” inhibitors from large chemical libraries. For comparison, consensus kNN QSAR models for DHFR, ACE and Cox2 inhibitors were developed by using the same sets of training compounds and Chembench with the parameters adjusted to reproduce the results of the published HDAC inhibitor QSAR screening studies.[137] For each target, the Chembench generated kNN QSAR models with the best performance against the 5-

fold CV testing compounds (**Table 4.4**) were used as a consensus kNN QSAR model for comparison with our consensus SVR model.

The R^2 values of our SVR QSAR models in the 5-fold CV tests are in the range of 0.51-0.81, 0.43-0.85 and 0.45-0.79 for DHFR, ACE and Cox2 inhibitors respectively, which are comparable to the R^2 values (0.55-0.60, 0.60-0.61 and 0.80-0.86) of the Chembench generated kNN QSAR models tested on the same sets of inhibitors and non-inhibitors, and close to the reported average R^2 values of the PSA-Docking (0.6-0.8)[130-132] and the BTFAP (0.6)[133] methods for kinase inhibitors that have been randomly divided into 75/25 training and testing sets. The 75/25 split is very similar to our 5-fold CV set-up. Hence, the results by both testing methods may be reasonably compared with each other. These R^2 values are significantly above the success criterion of 0.32 derived from an extensive study of docking methods.[289] Therefore, our method performed equally well in activity prediction as the established QSAR methods.[130-134, 137]

Table 4.1 The 5-fold cross validation performance of the top-15 SVR QSAR models for predicting DHFR inhibitors

Model No.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Epsilon		0.21	0.22	0.23	0.23	0.24	0.25	0.26	0.27	0.26	0.27	0.28	0.29	0.30	0.31	0.32
Sigma		0.21	0.21	0.21	0.22	0.22	0.22	0.22	0.22	0.23	0.23	0.23	0.23	0.23	0.23	0.23
Internal training r^2	1	0.8059	0.8068	0.8076	0.7929	0.7936	0.7940	0.7943	0.7949	0.7810	0.7816	0.7823	0.7831	0.7838	0.7843	0.7847
	2	0.7894	0.7905	0.7916	0.7754	0.7766	0.7775	0.7783	0.7791	0.7637	0.7645	0.7649	0.7655	0.7666	0.7673	0.7677
	3	0.7873	0.7875	0.7875	0.7744	0.7746	0.7747	0.7749	0.7751	0.7619	0.7623	0.7625	0.7627	0.7627	0.7629	0.7631
	4	0.7903	0.7906	0.7909	0.7800	0.7810	0.7813	0.7813	0.7815	0.7695	0.7703	0.7710	0.7714	0.7717	0.7719	0.7719
	5	0.7640	0.7652	0.7662	0.7524	0.7532	0.7541	0.7549	0.7557	0.7422	0.7428	0.7434	0.7440	0.7448	0.7460	0.7470
Internal testing r^2	1	0.5076	0.5079	0.5079	0.5071	0.5079	0.5085	0.5082	0.5074	0.5100	0.5088	0.5075	0.5072	0.5065	0.5054	0.5046
	2	0.5660	0.5655	0.5648	0.5592	0.5598	0.5599	0.5592	0.5580	0.5529	0.5537	0.5530	0.5520	0.5514	0.5503	0.5490
	3	0.5836	0.5875	0.5921	0.5796	0.5848	0.5892	0.5935	0.5967	0.5826	0.5890	0.5926	0.5947	0.5961	0.5961	0.5972
	4	0.5591	0.5592	0.5585	0.5528	0.5528	0.5529	0.5533	0.5539	0.5487	0.5491	0.5486	0.5479	0.5470	0.5468	0.5477
	5	0.5358	0.5384	0.5398	0.5407	0.5394	0.5381	0.5374	0.5376	0.5389	0.5387	0.5383	0.5372	0.5352	0.5360	0.5364
Predictive q^2		0.5278	0.5293	0.5304	0.5273	0.5284	0.5291	0.5296	0.5298	0.5272	0.5287	0.5287	0.5282	0.5272	0.5267	0.5267
External inactive accuracy	1	0.9980	0.9981	0.9980	0.9981	0.9981	0.9981	0.9981	0.9981	0.9983	0.9983	0.9982	0.9981	0.9981	0.9981	0.9981
	2	0.9982	0.9981	0.9981	0.9982	0.9981	0.9981	0.9981	0.9981	0.9982	0.9982	0.9982	0.9982	0.9981	0.9981	0.9981
	3	0.9983	0.9983	0.9983	0.9982	0.9982	0.9982	0.9982	0.9982	0.9983	0.9982	0.9982	0.9982	0.9983	0.9982	0.9982
	4	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9983	0.9982	0.9982	0.9981	0.9982	0.9982	0.9982
	5	0.9983	0.9983	0.9983	0.9985	0.9984	0.9983	0.9983	0.9982	0.9984	0.9983	0.9983	0.9983	0.9983	0.9983	0.9983
	Average	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9983	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982

Table 4.2 The 5-fold cross validation performance of the top-15 SVR QSAR models for predicting ACE inhibitors

Model No.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Epsilon		0.47	0.48	0.52	0.53	0.54	0.55	0.56	0.57	0.58	0.59	0.60	0.53	0.54	0.55	0.56
Sigma		0.15	0.15	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.14	0.14	0.14	0.14
Internal training r^2	1	0.8043	0.8039	0.8322	0.8313	0.8304	0.8249	0.8285	0.8277	0.8270	0.8262	0.8254	0.8151	0.8142	0.8133	0.8122
	2	0.8143	0.8139	0.8471	0.8459	0.8446	0.8434	0.8422	0.8410	0.8398	0.8386	0.8373	0.8282	0.8272	0.8262	0.8250
	3	0.8093	0.8089	0.8435	0.8425	0.8415	0.8404	0.8393	0.8381	0.8369	0.8356	0.8341	0.8240	0.8232	0.8224	0.8215
	4	0.8223	0.8218	0.8470	0.8456	0.8443	0.8428	0.8413	0.8397	0.8382	0.8367	0.8353	0.8336	0.8327	0.8317	0.8307
	5	0.8046	0.8043	0.8390	0.8376	0.8362	0.8348	0.8334	0.8321	0.8308	0.8293	0.8280	0.8209	0.8199	0.8190	0.8179
Internal testing r^2	1	0.5271	0.5268	0.5558	0.5581	0.5601	0.5614	0.5625	0.5631	0.5636	0.5640	0.5641	0.5404	0.5420	0.5433	0.5445
	2	0.4288	0.4286	0.4388	0.4396	0.4405	0.4415	0.4426	0.4436	0.4446	0.4454	0.4460	0.4327	0.4329	0.4331	0.4331
	3	0.5399	0.5400	0.5336	0.5333	0.5331	0.5328	0.5323	0.5320	0.5317	0.5312	0.5308	0.5333	0.5338	0.5342	0.5343
	4	0.4667	0.4664	0.4756	0.4758	0.4759	0.4764	0.4770	0.4772	0.4774	0.4777	0.4776	0.4709	0.4713	0.4722	0.4731
	5	0.5110	0.5110	0.5147	0.5151	0.5152	0.5151	0.5152	0.5151	0.5147	0.5140	0.5133	0.5104	0.5108	0.5107	0.5106
Predictive q^2		0.4557	0.4558	0.4553	0.4562	0.4569	0.4575	0.4580	0.4583	0.4584	0.4583	0.4579	0.4553	0.4561	0.4568	0.4572
External inactive accuracy	1	0.9934	0.9934	0.9940	0.9939	0.9938	0.9938	0.9938	0.9937	0.9936	0.9935	0.9935	0.9936	0.9935	0.9934	0.9934
	2	0.9941	0.9941	0.9940	0.9940	0.9939	0.9938	0.9938	0.9937	0.9937	0.9937	0.9937	0.9940	0.9939	0.9940	0.9938
	3	0.9930	0.9929	0.9938	0.9937	0.9937	0.9936	0.9935	0.9933	0.9932	0.9932	0.9932	0.9934	0.9934	0.9933	0.9933
	4	0.9944	0.9944	0.9945	0.9945	0.9944	0.9944	0.9943	0.9942	0.9941	0.9942	0.9941	0.9944	0.9943	0.9943	0.9942
	5	0.9938	0.9937	0.9939	0.9938	0.9938	0.9937	0.9938	0.9936	0.9934	0.9934	0.9933	0.9935	0.9934	0.9934	0.9934
	Average	0.9937	0.9937	0.9940	0.9940	0.9939	0.9938	0.9938	0.9937	0.9936	0.9936	0.9936	0.9938	0.9937	0.9937	0.9936

Table 4.3 The 5-fold cross validation performance of the top-15 SVR QSAR models for predicting Cox2 inhibitors

Model No.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Epsilon		0.19	0.20	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.34	0.35	0.36	0.37	0.38	0.39
Sigma		0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.17	0.17	0.17	0.17	0.17	0.17
Internal training r^2	1	0.7731	0.7743	0.7753	0.7762	0.7771	0.7777	0.7782	0.7786	0.7790	0.7508	0.7511	0.7513	0.7515	0.7518	0.7521
	2	0.7657	0.7667	0.7678	0.7689	0.7699	0.7708	0.7717	0.7725	0.7732	0.7438	0.7443	0.7448	0.7450	0.7452	0.7455
	3	0.7794	0.7806	0.7816	0.7826	0.7836	0.7845	0.7852	0.7858	0.7864	0.7562	0.7566	0.7569	0.7571	0.7571	0.7571
	4	0.7832	0.7843	0.7852	0.7859	0.7866	0.7872	0.7877	0.7882	0.7887	0.7598	07602	0.7607	0.7611	0.7615	0.7619
	5	0.7856	0.7866	0.7874	0.7880	0.7886	0.7891	0.7896	0.7901	0.7905	0.7606	0.7611	0.7616	0.7620	0.7624	0.7627
Internal testing r^2	1	0.5113	0.5114	0.5115	0.5114	0.5115	0.5112	0.5108	0.5102	0.5097	0.5038	0.5040	0.5043	0.5044	0.5043	0.5041
	2	0.4899	0.4902	0.4904	0.4907	0.4909	0.4909	0.4909	0.4908	0.4905	0.4896	0.4892	0.4887	0.4883	0.4878	0.4876
	3	0.4651	0.4657	0.4662	0.4668	0.4675	0.4679	0.4685	0.4691	0.4693	0.4674	0.4671	0.4666	0.4657	0.4649	0.4639
	4	0.4525	0.4519	0.4515	0.4509	0.4503	0.4500	0.4499	0.4498	0.4497	0.4522	0.4526	0.4528	0.4530	0.4529	0.4531
	5	0.4673	0.4670	0.4663	0.4653	0.4648	0.4645	0.4640	0.4631	0.4625	0.4573	0.4580	0.4588	0.4593	0.4598	0.4601
Predictive q^2		0.3771	0.3775	0.3777	0.3777	0.3779	0.3779	0.3780	0.3777	0.3774	0.3770	0.3774	0.3777	0.3776	0.3773	0.3770
External inactive accuracy	1	0.9682	0.9680	0.9679	0.9676	0.9676	0.9676	0.9677	0.9676	0.9672	0.9673	0.9669	0.9668	0.9665	0.9665	0.9663
	2	0.9671	0.9670	0.9669	0.9668	0.9666	0.9666	0.9663	0.9662	0.9659	0.9664	0.9663	0.9661	0.9661	0.9660	0.9658
	3	0.9677	0.9675	0.9675	0.9673	0.9670	0.9670	0.9669	0.9667	0.9666	0.9682	0.9680	0.9679	0.9677	0.9675	0.9674
	4	0.9678	0.9676	0.9676	0.9676	0.9673	0.9671	0.9671	0.9671	0.9669	0.9678	0.9677	0.9675	0.9673	0.9671	0.9669
	5	0.9675	0.9688	0.9675	0.9674	0.9676	0.9677	0.9675	0.9670	0.9668	0.9671	0.9672	0.9672	0.9669	0.9665	0.9663
	Average	0.9677	0.9676	0.9675	0.9673	0.9672	0.9672	0.9671	0.9669	0.9667	0.9673	0.9672	0.9671	0.9669	0.9667	0.9665

Table 4.4 The performance of SVR and Chembench kNN QSAR in predicting the activity of DHFR, ACE and Cox2 inhibitors within and outside similarity-based applicability domain (AD)

	DHFR inhibitors		**ACE inhibitors**		**Cox2 inhibitors**	
	SVR	Chembench kNN	SVR	Chembench kNN	SVR	Chembench kNN
R^2 in 5-fold cross-validation tests	0.51-0.60	0.55-0.59	0.43-0.56	0.60-0.61	0.45-0.49	0.80-0.84
R^2 for post-2010 compounds within similarity-based AD	0.32	0.31	0.32	0.15	0.19	0.15
R^2 for post-2010 compounds outside similarity-based AD	NA	NA	0.26	NA	0.15	NA

The ability of our models in predicting “new” inhibitors within the similarity-based ADs was tested by using the 26, 41 and 59 post-2010 DHFR, ACE and Cox2 inhibitors and 42, 20 and 43 post-2010 non-inhibitors inside the similarity-based ADs defined by the method of the Tropsha group.[134, 137] **Figure 4.4-4.6** show the comparison of the actual and the predicted pIC_{50} values of our models and those of the Chembench generated consensus kNN QSAR models in identifying these “new” compounds. The R^2 values of our models for predicting these “new” DHFR, ACE and Cox2 inhibitors and non-inhibitors are 0.32, 0.32 and 0.19 respectively, which are comparable to (or slightly better than) those of the consensus kNN QSAR models. These R^2 values are substantially lower than those evaluated by the 5-fold CV, but nonetheless close to the success criterion of 0.32.[289] One possible reason for the lower R^2 values is the higher level of structural novelty of the post-2010 vs the pre-2010 compounds than the training vs testing compounds in a 5-fold CV setting. Additionally, the relatively lower R^2 for predicting “new” Cox2 inhibitors may further be attributed to the much higher diversity of the collected Cox2 inhibitors included in the training set (spread in 901 families vs. 76 families for DHFR and 188 families for ACE).

The performance of our models in predicting highly novel inhibitors outside similarity-based ADs was assessed by using the 0, 6 and 13 post-2010 DHFR, ACE and Cox2 inhibitors and 4, 34 and 7 post-2010 non-inhibitors outside the similarity-based ADs defined by the method of the Tropsha group.[134, 137] The comparison of the actual and the predicted pIC_{50} values of our models in identifying these highly novel compounds is also shown in **Figure 4.4-4.6**. The R^2 values of our models for predicting these highly novel DHFR, ACE and Cox2 inhibitors and non-inhibitors are 0.04, 0.26 and 0.15 respectively, which are slightly lower than those in predicting “new” inhibitors inside the similarity-based ADs. Therefore, our method has some level of capability in predicting the activity of highly novel actives outside similarity-based ADs.

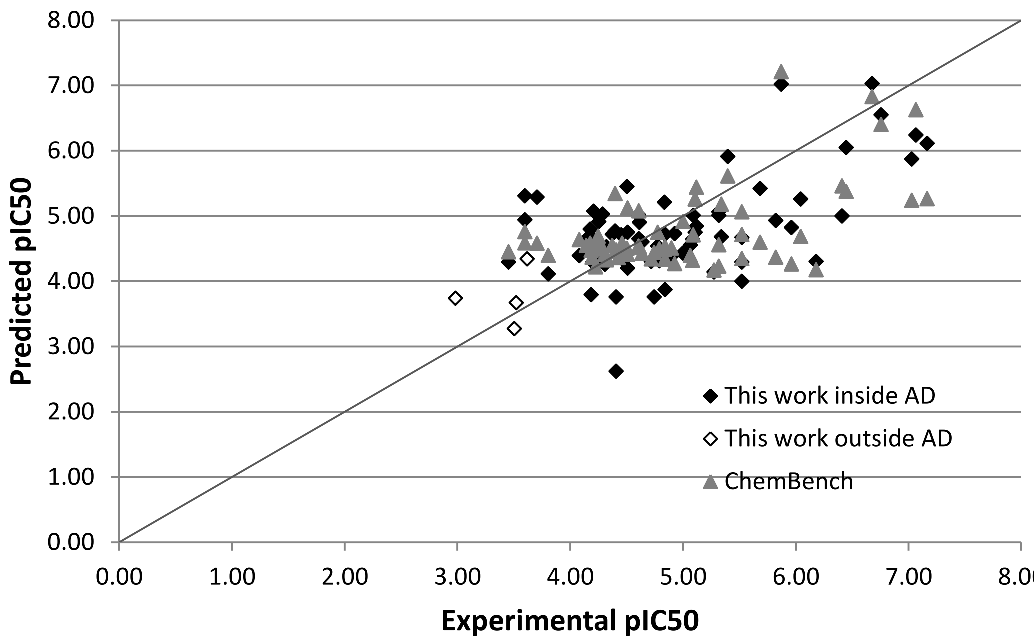


Figure 4.4 The comparison of the actual and the predicted pIC50 values of SVR and ChemBench kNN QSAR models trained by pre-2010 inhibitors in predicting the activity of post-2010 DHFR inhibitors and non-inhibitors inside and outside similarity-based applicability domain (AD)

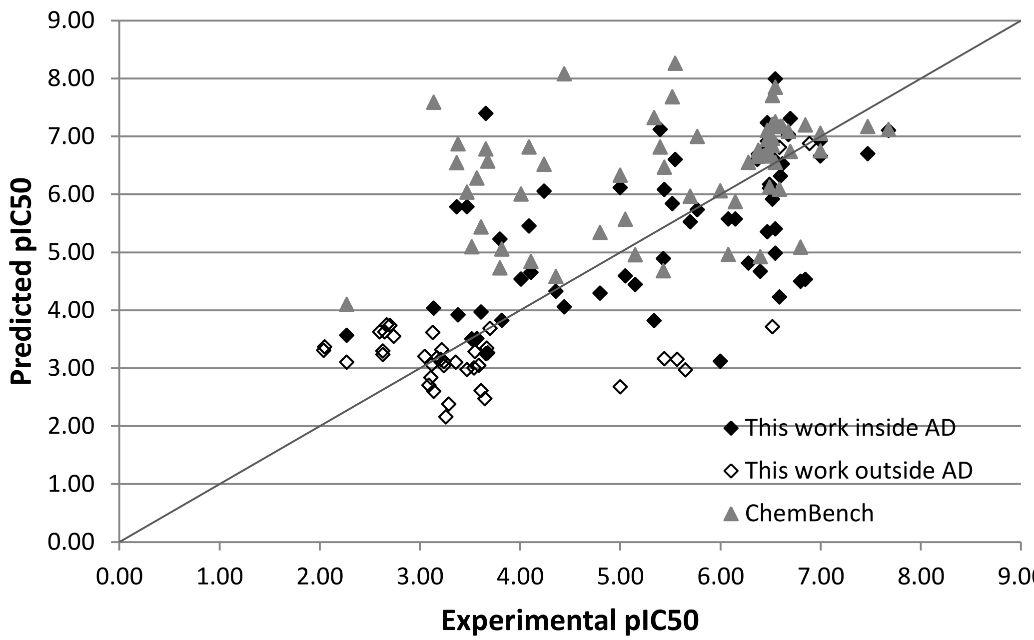


Figure 4.5 The comparison of the actual and the predicted pIC50 values of SVR and ChemBench kNN QSAR models trained by pre-2010 inhibitors in predicting the activity of post-2010 ACE inhibitors and non-inhibitors inside and outside similarity-based applicability domain (AD)

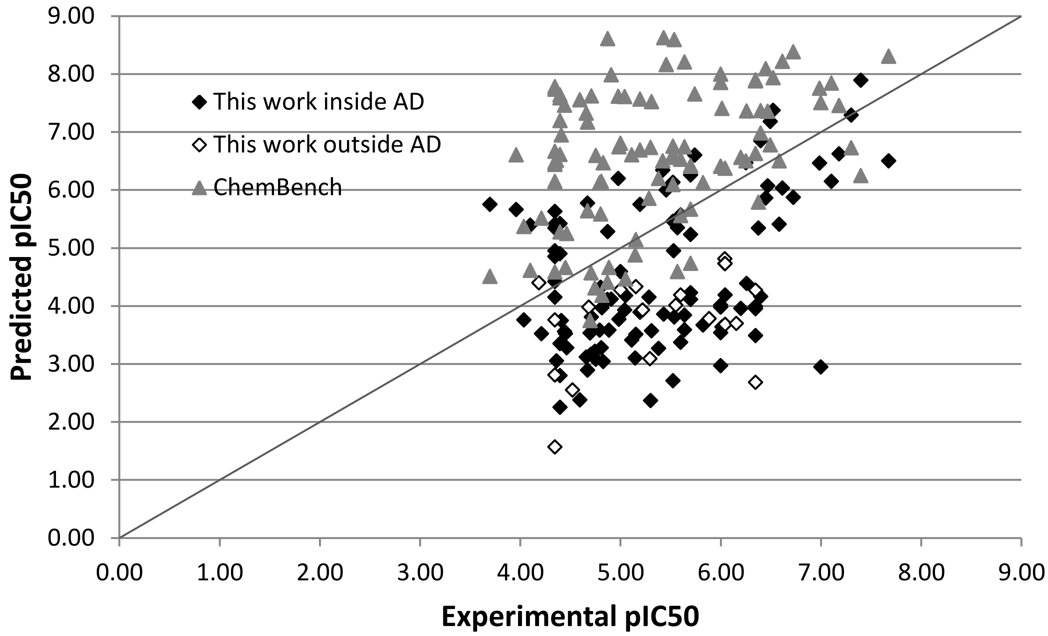


Figure 4.6 The comparison of the actual and the predicted pIC50 values of SVR and ChemBench kNN QSAR models trained by pre-2010 inhibitors in predicting the activity of post-2010 Cox2 inhibitors and non-inhibitors inside and outside similarity-based applicability domain (AD)

4.3.2 Virtual screening performance of SVR QSAR models in searching DHFR, ACE and Cox2 inhibitors from large libraries

In evaluating the VS performance of our models in screening large chemical libraries, we used our models and the Chembench generated consensus kNN QSAR models to screen 168K MDDR compounds for identifying the 142, 357 and 939 known DHFR, ACE and Cox2 patented inhibitors that are inside the similarity-based ADs defined by the method of the Tropsha group[134, 137] (**Table 4.5)**. A compound was identified as a virtual hit if the predicted pIC_{50}>5. VS performance is typically measured by three quantities: yield (ratio of the identified and all known inhibitors in the searched libraries), hit rate (ratio of the identified inhibitors and all virtual hits) and enrichment factor (ratio of hit rate and random selection rate, which measures improvement over random selection). The yield, hit rate and enrichment factor of the DHFR, ACE and Cox2 SVR QSAR models are 85.2%, 34.6% and 409.0 for DHFR, 86.3%, 30.5% and 143.4 for ACE, and 71.0%, 26.0% and 46.5 for Cox2 respectively, which are comparable to those of 81.0%, 21.1% and 249.2 for DHFR, 88.2%, 11.5% and 54.1 for ACE, and 66.9%, 9.2% and 16.5 for Cox2 by the Chembench kNN QSAR models. These results suggest that our method is capable of searching large chemical libraries at comparable yield and substantially improved hit rate and enrichment factor with respect to such established methods as the Chembench generated consensus kNN QSAR models.

We further evaluated the capability of our models in searching highly novel actives from large chemical libraries by screening 168K MDDR compounds for identifying the 25, 175 and 51 known DHFR, ACE and Cox2 patented inhibitors that are outside the similarity-based ADs defined by the method of the Tropsha group[134, 137] (**Table 4.5**). The yield, hit rate, and enrichment factor of our models in identifying these highly novel DHFR, ACE and Cox2 from 168K MDDR compounds are 40.0%, 2.3% and 152.7 for DHFR, 45.7%, 1.7% and 16.1 for ACE, and 19.6%, 0.18% and 5.9 for Cox2 respectively, which suggests that our method has some level of capability in finding highly novel actives from large chemical libraries. Moreover, the VS performance our models in searching large chemical libraries were tested by screening 13.56 million PubChem compounds, which identified 26,217 (0.19%), 122,829 (0.91%) and 559,279 (4.12%) of the PubChem compounds as virtual DHFR, ACE and Cox2 inhibitor hits respectively. Even if all of these virtual hits turn out to be false, the maximum false hit rate would be no more than 0.19%, 0.91% and 4.12% respectively. Therefore, our method is capable of searching large chemical libraries at very low false hit rate. We also analyzed the similarity levels of our identified 26,217, 122,829 and 559,279 PubChem virtual DHFR, ACE and Cox2 inhibitor hits with respect to the pre-2010 DHFR, ACE and Cox2 inhibitors, which showed that these virtual hits are roughly equally distributed in different similarity ranges (**Table 4.6-4.8** and **Figure 4.7-4.9**). This suggests that our QSAR models selected virtual hits not based on some form of similarity but rather based on the differentiating features derived from the known pre-2010 inhibitors and the putative non-inhibitors.

Table 4.5 The performance of SVR and ChemBench kNN QSAR models trained by the same sets of pre-2010 inhibitors in searching 168K MDDR compounds for identifying the 167, 532 and 990 patented DHFR, ACE and Cox2 inhibitors within and outside similarity-based applicability domain (AD)

		DHFR inhibitors		**ACE inhibitors**		**Cox2 inhibitors**	
		SVR	Chembench kNN	SVR	Chembench kNN	SVR	Chembench kNN
Within similarity-based AD	No of compounds	3,685		3,706		15,842	
	No of patented inhibitors	142		357		939	
	No of virtual hits	350	546	1,011	2,739	2,566	6,800
	No of patented inhibitors identified	121	115	308	315	667	628
	Yield	85.2%	81.0%	86.3%	88.2%	71.0%	66.9%
	Hit rate	34.6%	21.1%	30.5%	11.5%	26.0%	9.2%
	Enrichment factor	409.0	249.2	143.4	54.1	46.5	16.5
Outside similarity-based AD	No of compounds	164,295		164,283		152,109	
	No of patented inhibitors	25		175		51	
	No of virtual hits	440	NA	4,767	NA	5,561	NA
	No of patented inhibitors identified	10	NA	80	NA	10	NA
	Yield	40.0%	NA	45.7%	NA	19.6%	NA
	Hit rate	2.3%	NA	1.7%	NA	0.18%	NA
	Enrichment factor	152.7	NA	16.1	NA	5.9	NA

Table 4.6 The similarity levels of our identified PubChem virtual DHFR, inhibitor hits with respect to the pre-2010 DHFR inhibitors

Tanimoto	Total	<10uM	%<10uM	<1uM	%<1uM	<100nM	%<100nM
0.0-0.1	168856	224	0.1327%	57	0.0338%	16	0.0095%
0.1-0.2	541160	517	0.0955%	152	0.0281%	76	0.0140%
0.2-0.3	1460146	1828	0.1252%	543	0.0372%	227	0.0155%
0.3-0.4	1196544	2993	0.2501%	1173	0.0980%	295	0.0247%
0.4-0.5	1688772	2292	0.1357%	479	0.0284%	106	0.0063%
0.5-0.6	3420570	7940	0.2321%	2574	0.0753%	686	0.0201%
0.6-0.7	3734182	7158	0.1917%	2188	0.0586%	799	0.0214%
0.7-0.8	1157503	3192	0.2758%	1190	0.1028%	238	0.0206%
0.8-0.9	180968	73	0.0403%	4	0.0022%	1	0.0006%
0.9-1.0	12019	0	0.0000%	0	0.0000%	0	0.0000%
Total	13560720	26217	0.1933%	8360	0.0616%	2444	0.0180%

Table 4.7 The similarity levels of our identified PubChem virtual ACE, inhibitor hits with respect to the pre-2010 ACE inhibitors

Tanimoto	Total	<10uM	%<10uM	<1uM	%<1uM	<100nM	%<100nM
0.0-0.1	122682	1259	1.0262%	228	0.1858%	25	0.0204%
0.1-0.2	93791	517	0.5512%	72	0.0768%	17	0.0181%
0.2-0.3	450593	5792	1.2854%	1453	0.3225%	431	0.0957%
0.3-0.4	937720	7728	0.8241%	1947	0.2076%	777	0.0829%
0.4-0.5	1516315	11707	0.7721%	2493	0.1644%	759	0.0501%
0.5-0.6	2889486	25333	0.8767%	5394	0.1867%	1480	0.0512%
0.6-0.7	5051559	46274	0.9160%	10646	0.2107%	3028	0.0599%
0.7-0.8	2160888	21128	0.9777%	4722	0.2185%	1570	0.0727%
0.8-0.9	324139	2980	0.9194%	305	0.0941%	61	0.0188%
0.9-1.0	13547	111	0.8194%	15	0.1107%	3	0.0221%
Total	13560720	122829	0.9058%	27275	0.2011%	8151	0.0601%

Table 4.8 The similarity levels of our identified PubChem virtual Cox2, inhibitor hits with respect to the pre-2010 Cox2 inhibitors

Tanimoto	Total	<10uM	%<10uM	<1uM	%<1uM	<100nM	%<100nM
0.0-0.1	111621	4716	4.2250%	909	0.8144%	140	0.1254%
0.1-0.2	80528	3006	3.7329%	844	1.0481%	191	0.2372%
0.2-0.3	332286	12928	3.8906%	2763	0.8315%	797	0.2399%
0.3-0.4	793811	26527	3.3417%	4247	0.5350%	708	0.0892%
0.4-0.5	958893	40430	4.2163%	8417	0.8778%	2133	0.2224%
0.5-0.6	1090704	47289	4.3356%	9988	0.9157%	1840	0.1687%
0.6-0.7	3659030	141041	3.8546%	26519	0.7248%	5541	0.1514%
0.7-0.8	5067798	224226	4.4245%	47168	0.9307%	12078	0.2383%
0.8-0.9	1385021	57036	4.1181%	10501	0.7582%	2352	0.1698%
0.9-1.0	81028	2080	2.5670%	506	0.6245%	144	0.1777%
Total	13560720	559279	4.1243%	111862	0.8249%	25924	0.1912%

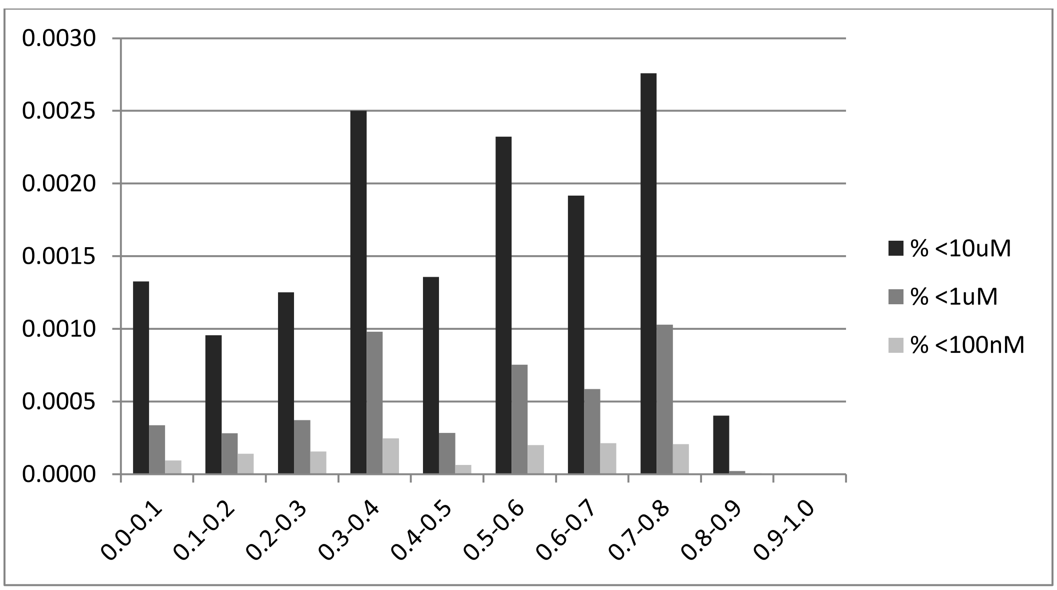


Figure 4.7 The similarity levels of our identified PubChem virtual DHFR inhibitor hits with respect to the pre-2010 DHFR inhibitors

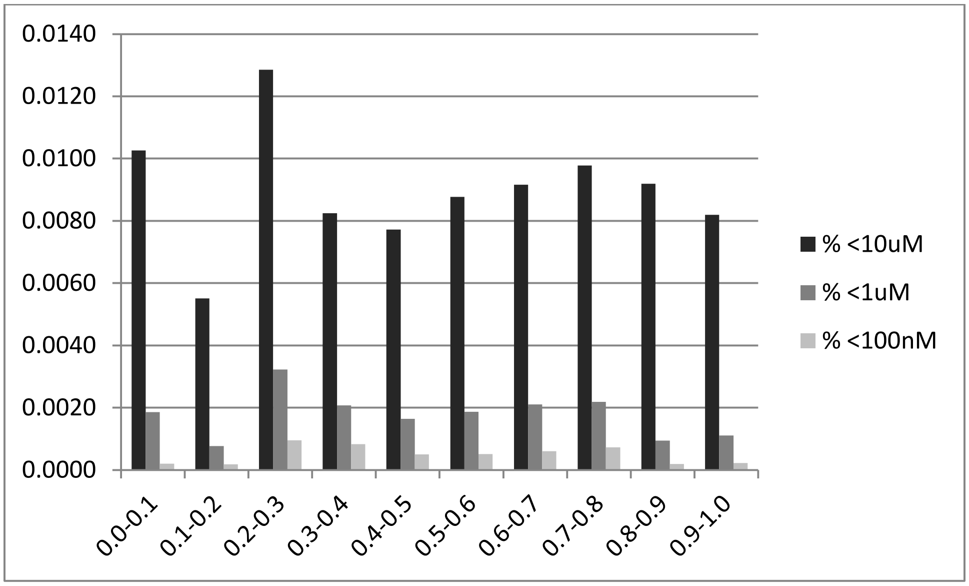


Figure 4.8 The similarity levels of our identified PubChem virtual ACE inhibitor hits with respect to the pre-2010 ACE inhibitors

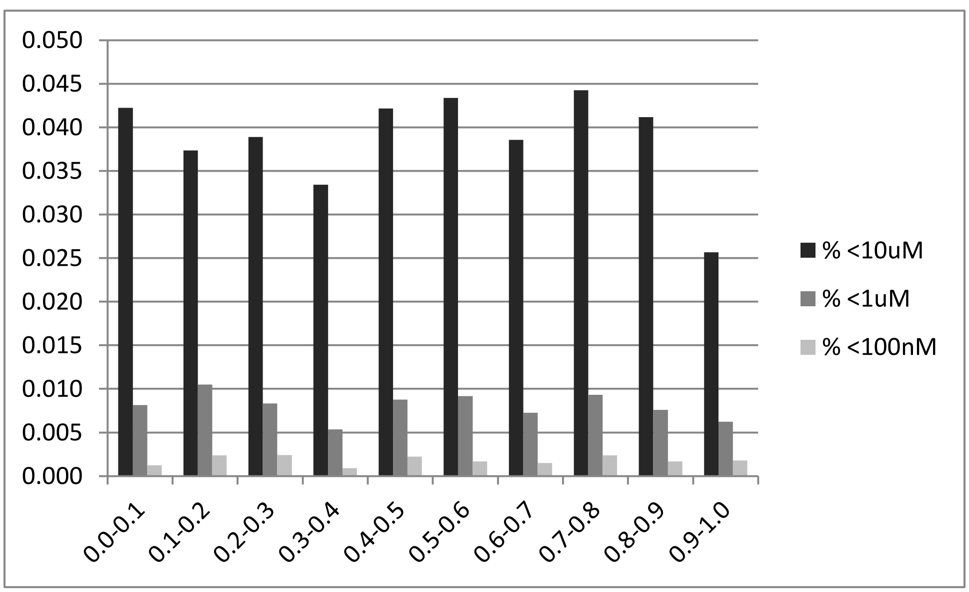


Figure 4.9 The similarity levels of our identified PubChem virtual Cox2 inhibitor hits with respect to the pre-2010 Cox2 inhibitors

CHAPTER 5 Virtual Screening of Selective Multi-target Kinase Inhibitors

As illustrated in **Chapter 3,** one potential application of the Pathway Cross-talk Database (PCD) lies in facilitating system level studies of diseases and mechanism of drug combinations which was demonstrated by the analysis of the effect of glutamate on glioma cell invasion and the synergistic actions of tamoxifen-herceptin drug combination. Another potential usage of PCD is the systematic analysis of target combinations regulating multiple disease-related signaling pathways thus facilitating the discovery of multi-target agents.

Multi-target agents have been increasingly explored for enhancing efficacy and reducing counter-target activities and toxicities. Efficient virtual screening (VS) tools for searching selective multi-target agents are desired. In **Chapter 4,** an epsilon-Support Vector Regression (ε-SVR) based high-throughput QSAR approach was developed and tested on DHFR, ACE and Cox2 inhibitors. In this chapter, this approach is applied as the VS tool for searching dual-inhibitors of 4 combinations of 5 anticancer kinase targets, EGFR, VEGFR, PDGFR, Src and FGFR.

5.1 Introduction

Large percentage of drugs in development, which are typically directed at an individual target, frequently show reduced efficacies and undesired safety and resistance profiles due to network robustness,[17] redundancy,[290] cross-talk,[225] compensatory and neutralizing actions,[291] anti-target and counter-target activities,[292] and on-target and off-target toxicities.[293] Multi-target agents and drug-combinations have been increasingly explored[16, 17] for enhancing therapeutic efficacies and improving safety and resistance profiles by selectively modulating the elements of these counter-target and toxicity activities.[18] In particular, multi-target kinase inhibitors are among the most successful clinical anticancer drugs (e.g. sunitinib against PDGFR and VEGFR, dasatinib against Abl and Src, sorafenib against Braf and VEGFR, and lapatinib against EGFR and HER2) and have been actively pursued in current drug discovery efforts.[28, 294] Methods for efficient search of multi-target agents are highly desired.

Virtual screening (VS) methods have been widely explored for facilitating lead discovery against individual targets.[276, 295, 296] In particular, molecular docking,[80] pharmacophore,[297] QSAR,[298] machine learning,[299] and combination methods[300] have been extensively used for VS of single-target kinase inhibitors, but few multi-target VS studies have been reported.[301, 302] An interesting strategy for identifying multi-target kinase inhibitors is to use experimentally obtained small-scale profiles for predicting inhibitors of a larger kinase set.[302] In principle, single-target VS tools may be combined to collectively identify multi-target agents, which is practically useful if the individual VS tools have sufficiently high yields and low false-hit rates. High yields compensate for the reduced collective yields of combinatorial VS tools (For two statistically-independent VS tools of 50%-70% yields, the collective yield of their combination is roughly the product of the yield of individual tools, which is 25%-49%). Low false-hit rates are needed for high enrichment factors in searching multi-target agents that are significantly fewer in numbers and more sparsely distributed in the chemical space than non-dual inhibitors (**Table 5.1**).

A support vector regression (SVR) based high throughput QSAR method has been developed and may be potentially explored as multi-target VS tools because it has shown high yields and low false-hit rates in searching single-target agents for DHFR, ACE and Cox2 and is able to identify highly novel inhibitors even outside the similarity-based ADs. This method identifies active compounds in fast-speed by differentiating physicochemical profiles rather than structural similarity to active compounds *per se*, and requires no knowledge of target structure and no computation of structural flexibility, activity-related features, solvation effects and binding affinities. The multi-target VS performance of this SVR QSAR method, which combine the prediction of two separate SVR QSAR models for each the multiple kinases, was tested by using it to search dual-inhibitors of combinations of 5 anticancer kinase targets EGFR, VEGFR, PDGFR, FGFR and Src. **Figure 5.1** shows the illustration of using SVR QSAR methods for searching multi-target inhibitors. These kinase targets were selected because of their therapeutic relevance and the availability of sufficient number of the known inhibitors and dual-inhibitors.

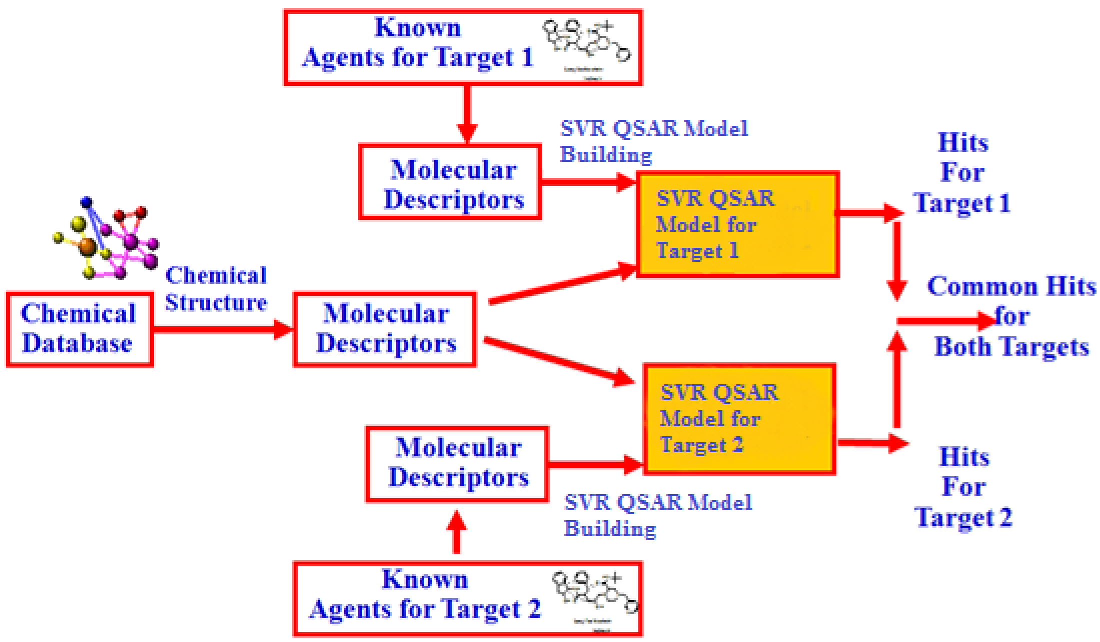


Figure 5.1 Illustration of using SVR QSAR method for searching multi-target inhibitors

Based on dual-inhibitor availability, we focused on 4 kinase-pairs EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src. As described in **Section 3.3.3,** these kinase-pairs are frequently co-expressed or co-activated in various cancers e.g. NSCLC,[303, 304] and targeted by multi-target agents[28, 294] with good anticancer efficacies. Inhibitors of growth factor receptor tyrosine kinases EGFR, VEGFR, PDGFR and FGFR have been successfully used for cancer treatments,[28, 305-309] EGFR promotes proliferation and survival.[305] VEGFR regulates angiogenesis and survival[307]. PDGFR modulates angiogenesis and growth, and is one of the multi-targets of several approved and clinical trial drugs.[28, 308] FGFR regulates angiogenesis and cancer progression, and is one of the multi-targets of several clinical trial drugs.[28, 309] Src modulates multiple pathways of cell growth, differentiation, migration and survival, and is part of the multi-targets of several marketed and clinical trial drugs.[28, 310]

Multi-target VS performance was tested by a rigorous method that assumes there is no explicit knowledge of known multi-target agents, because the number of known multi-target agents are generally small for many target-pairs. SVR QSAR models of each kinase were trained by using non-dual inhibitors of that kinase. The collective yield of SVR QSAR models of each kinase-pair (percent of known dual-inhibitors identified as dual-inhibitors) was estimated by using known dual-inhibitors of each kinase-pair. Target selectivity of each SVR QSAR model was assessed by using non-dual inhibitors of the kinase-pair and inhibitors of the other 3 kinases, out of the 5 evaluated kinases, not included in the kinase-pair. Virtual-hit rates and false-hit rates in searching large compound libraries were evaluated by using 13.56 million PubChem, 168 thousand compounds from the MDL Drug Data Report (MDDR) database, and 1,175-9,356 MDDR compounds similar in structural and physicochemical properties to the known dual-kinase inhibitors. MDDR contains biologically relevant compounds (active against individual molecular target or biological assay) and well-defined derivatives reported in the patent literature, journals, meetings and congresses. PubChem and MDDR contain high percentages of inactive or active compounds significantly different from the dual-inhibitors, and the easily distinguishable features may make VS enrichments artificially good.[311] Therefore, VS performance is more strictly tested by using subset of MDDR compounds similar to the dual-inhibitors so that enrichment is not simply a separation of trivial physicochemical features.[219]

5.2 Materials and methods

5.2.1 Compound collection, training and testing datasets, molecular descriptors

A total of 428-2,912 non-dual inhibitors of EGFR, VEGFR, PDGFR, FGFR and Src, and 67-256 dual inhibitors of EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src, each with $IC_{50} \leq 10\mu M$, were collected from ChEMBL database[158]. Dual-inhibitors and non-dual inhibitors of a kinase-pair refer to inhibitors of both and one of the two kinases respectively regardless of their activities against other kinases. **Table 5.1** summarizes the statistics of these inhibitors and MDDR compounds similar to at least one of the dual-inhibitors. **Figure 5.2** shows the Venn graph of our collected dual-inhibitors the 4 evaluated kinase pairs and non-dual-inhibitors of the 5 evaluated kinases. As few non-inhibitors have been reported, putative non-inhibitors of each kinase were generated by following the same protocol as described previously in **Section 2.2.1.3** and **Section 4.2.1**. As a result, a total of 7,628-8,241 compounds extracted from the 7,628-8,241 families (1 per family) that contain no known inhibitor were used as the putative non-inhibitors.

Table 5.1 Datasets of dual-inhibitors and non-dual-inhibitors of the kinase-pairs used for developing and testing combinatorial SVM dual-inhibitor virtual screening tools. Additional sets of 13.56 million PubChem compounds and 168 thousand MDDR active compounds were also used for the test.

Kinase pair	Kinase A – Kinase B		EGFR-VEGFR	EGFR-PDGFR	EGFR-FGFR	EGFR-Src
Inhibitors in training sets	Training set for Kinase A	No of inhibitors of A that are non-inhibitor of B (No of families)	2,142 (635)	2,343 (666)	2,308 (658)	2,150 (631)
		No of these inhibitors that are in the B inhibitor families (No of families)	1,309 (255)	457 (95)	455 (87)	672 (165)
		No of these inhibitors that are in the families of dual inhibitors (No of families)	600 (79)	217 (26)	244 (38)	368 (74)
	Training set for Kinase B	No of inhibitors of B that are non-inhibitor of A (No of families)	2,912 (795)	675 (212)	428 (182)	1,444 (437)
		No of these inhibitors that are in the A inhibitors families (No of families)	1,293 (255)	347 (95)	256 (87)	768 (165)
		No of these inhibitors that are in the families of dual inhibitors of A and B (No of families)	539 (83)	162 (16)	145 (35)	450 (72)
Inhibitors and Other Compounds in Testing Set	Dual Inhibitors of A and B	No of dual inhibitors of A and B (No of families)	256 (121)	67 (40)	91 (58)	256 (123)
		No (%) of dual inhibitors in the families that contain both A and B non-dual inhibitor in training sets	171 (66.8%)	28 (41.8%)	42 (46.2%)	122 (47.7%)
		No (%) of dual inhibitors of A and B as inhibitor of at least one of the other 3 kinases studied in this work	171 (66.8%)	45 (67.2%)	67 (73.6%)	146 (57.0%)
		No (%) of dual-inhibitors of A and B as inhibitor of more than 1 of the other 3 kinases studied in this work	21 (8.2%)	23 (34.3%)	23 (25.3%)	18 (7.0%)
	Inhibitors of other 3 kinases	No of inhibitors	1,816	4,051	4,298	3,282
	MDDR Compounds Similar to Dual Inhibitors of A and B	No of compounds	9,356	1,175	1,285	5,404

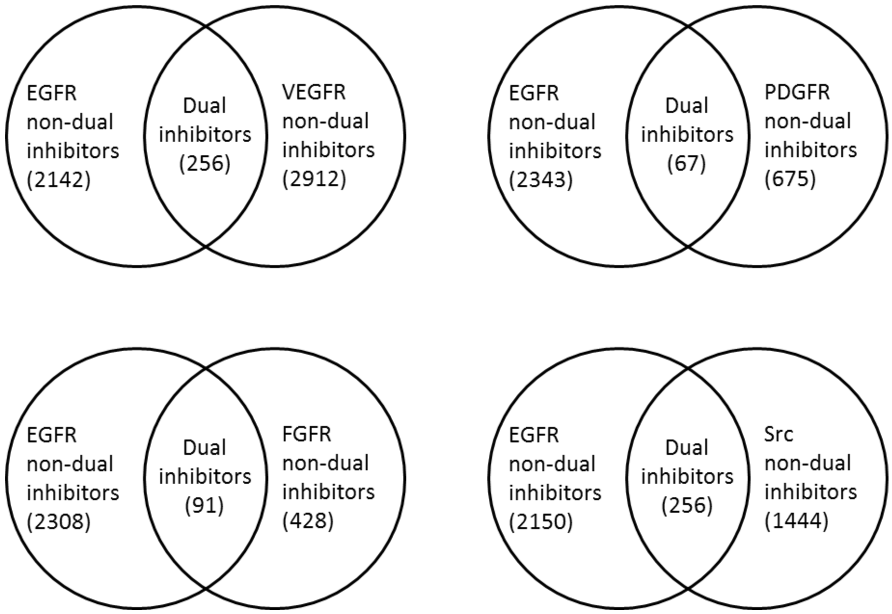


Figure 5.2 The Venn graph of the collected dual-inhibitors the 4 evaluated kinase-pairs and non-dual-inhibitors of the 5 evaluated kinases

The collected non-dual and dual inhibitors of EGFR, VEGFR, PDGFR, FGFR and Src, are distributed in 682, 833, 236, 205 and 488 families respectively, which is consistent with reported 191 unique scaffolds (154 clusters and 43 singletons) for 565 kinase inhibitors[299]. Because of the extensive efforts in searching kinase inhibitors, the number of undiscovered "inhibitor" families for each kinase in PubChem and MDDR is expected to be relatively small, most likely no more than several hundred families. The ratio of the "inhibitor" and "inactive" families for each kinase (hundreds families vs 7,628-8,241 families contained in PubChem and MDDR at present) is expected to be no more than ~999/8500, which is <13%. Therefore, putative non-inhibitor training dataset can be generated by extracting a few representative compounds from each of the families that contain no known inhibitor, with a maximum possible "wrong" prediction rate of <13% even in the extreme and unlikely cases that all of the undiscovered inhibitors are misplaced into the non-inhibitor class. The noise level generated by up to 13% "wrong" negative family representation is expected to be substantially smaller than the maximum 50% false-negative noise level tolerated by SVR QSAR models[312]. It is noted that 18.2%-25.0% of the dual-inhibitor families contain no non-dual inhibitor of the same kinase-pair, whose representative compounds were included in the inactive training datasets as dual-inhibitors are supposed to be unknown in our study. A substantial percentage of the dual-inhibitors in these "non-inhibitor" families were nonetheless identified as dual-inhibitors by our SVR QSAR models.

In this work, a total of 98 2D physicochemical descriptors generated from the MODEL[194] program were used. The detailed description of these molecular descriptors can be found in **Section 2.2.2**.

5.2.2 Computational models

A MLR method, support vector regression (SVR), is used for deriving QSAR models because it has consistently performed well,[135, 280-285] is less penalized by sample redundancy and has lower over-fitting risks.[286, 287] The objective of SVR is to find a function that minimally deviates from the activity values of the training compounds within a tube of radius ε.[211] For modeling chemically-diverse compounds, SVR typically maps the compounds into a higher dimensional space by using a kernel function. The detailed mathematical algorithms of SVR were described in **Section 2.2.3**. In this work, our SVR models were developed by using LIBSVM[288] with RBF kernel,[214-216] a hard margin C=1,000.

5.3 Results and discussion

5.3.1 Dual-inhibitors and non-dual inhibitors of the studied kinase-pairs

As shown in **Table 5.1**, the numbers of dual-inhibitors and non-dual inhibitors of the kinase-pairs are 256, 2,142 and 2,912 for EGFR-VEGFR, 67, 2,343 and 675 for EGFR-PDGFR, 91, 2,308 and 428 for EGFR-FGFR, and 256, 2,150 and 1,444 for EGFR-Src respectively. The dual-inhibitors and non-dual inhibitors are distributed in 40-123 and 182-795 families respectively. Hence, both the numbers and diversity of non-dual inhibitors and dual-inhibitors are at reasonable levels for developing and testing VS tools. The percentages of dual-inhibitors outside the common families of the non-dual inhibitors in the training datasets are 33.2% for EGFR-VEGFR, 58.2% for EGFR-PDGFR, 53.8% for EGFR-FGFR, and 52.3% for EGFR-Src respectively. Therefore, these dual-inhibitors have substantial degree of novelty against non-dual inhibitors. Moreover, 57.0%-73.6% of the dual-inhibitors of the kinase pairs are inhibitor of at least one of the other 3 kinases, but only up to 34.3% of the dual-inhibitors are inhibitor of at least 2 of the other 3 kinases. Hence, most of these dual-inhibitors are non-ubiquitous inhibitors and show some degree of kinase selectivity even though the majority of them target more than 2 kinases.

5.3.2 Virtual screening performance of SVR QSAR models in searching kinase dual-inhibitors from large libraries

The VS performance of SVR QSAR models in identifying dual-inhibitors of the 4 kinase-pairs is summarized in **Table 5.2** and further shown in **Figure 5.3**. The parameters of the developed SVR regression models for the evaluated kinases are in the ranges of ε=0.39-0.90 and σ=0.18-0.23. The dual-inhibitor yields are 42.2% for EGFR-VEGFR, 32.8% for EGFR-PDGFR, 22.0% for EGFR-FGFR, and 30.1% for EGFR-Src respectively. The yields for the intra- kinase pairs are comparable to the expected 25%-49% yields of combinations of good VS tools with individual yields of 50%-70%. Therefore, our SVR QSAR methods show reasonably good capability in identifying multi-target agents for kinase-pairs within a protein kinase group without requiring explicit knowledge of multi-target agents.

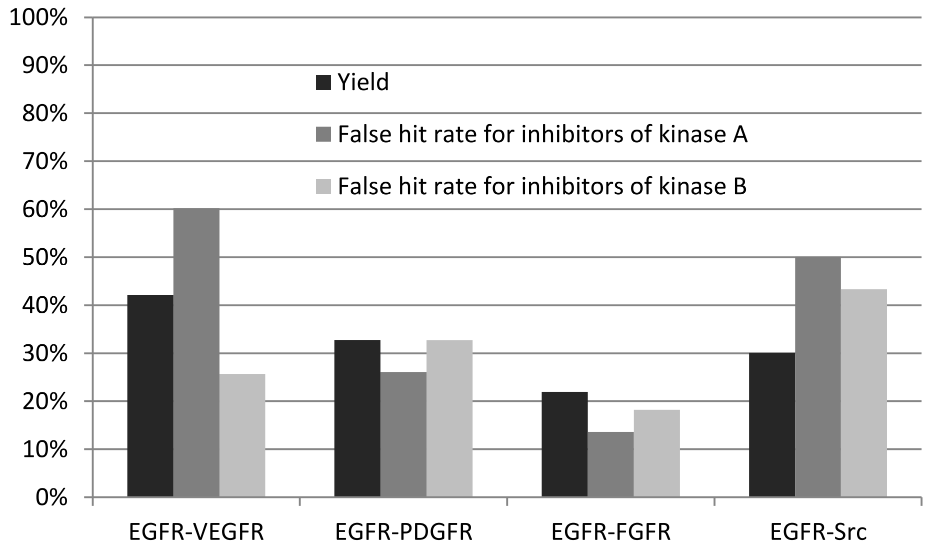


Figure 5.3 The VS performance of SVR QSAR models in identifying dual-inhibitors of 4 combinations of EGFR, VEGFR, PDGFR, FGFR and Src

Table 5.2 Virtual screening performance of SVR QSAR models for identifying dual-inhibitors of 4 combinations of EGFR, VEGFR, PDGFR, FGFR and Src

Kinase pair		EGFR-VEGFR	EGFR-PDGFR	EGFR-FGFR	EGFR-Src
Dual inhibitors	Yield (No of virtual hits)	42.2% (108)	32.8% (22)	22.0% (20)	30.1% (77)
	No (%) of identified true hits outside the common training active families of both kinases	27 (25.0%)	9 (40.9%)	9 (45.0%)	33 (42.9%)
Non-dual inhibitors of the same kinase pair	False hit rate for inhibitors of kinase A	60.2%	26.1%	13.6%	50.1%
	False hit rate for inhibitors of kinase B	25.7%	32.7%	18.2%	43.3%
Inhibitors of other 3 kinases	False hit rate	33.9%	18.3%	7.1%	12.5%
MDDR compounds similar to dual inhibitors	Virtual hit rate (No of virtual hits)	7.39% (691)	16.51% (194)	6.61% (85)	5.90% (319)
All 168 thousand MDDR compounds	Virtual hit rate (No of virtual hits)	1.55% (2,605)	0.93% (1,557)	0.39% (654)	0.99% (1,656)
13.56 million PubChem comnds	Virtual hit rate (No of virtual hits)	0.74% (102,497)	0.46% (61,764)	0.14% (18,981)	0.39% (52,498)

Target selectivity was tested by using SVR QSAR models to screen the 428-2,912 non-dual inhibitors of the 4 kinase-pairs, which misidentified 60.2% and 25.7% of the non-dual inhibitors of the kinase pair as dual-inhibitors for EGFR-VEGFR, 26.1% and 32.7% for EGFR-PDGFR, 13.6% and 18.2% for EGFR-FGFR, and 50.1% and 43.3% for EGFR-Src respectively. Therefore, these SVR QSAR models showed some selectivity in distinguishing dual-inhibitors from non-dual inhibitors yet with unsatisfactory false hit rate in some cases, e.g. 60.2% of the EGFR non-dual inhibitors were identified as EGFR-VEGFR dual inhibitors, and 50.1% of the EGFR non-dual inhibitors were identified as EGFR-Src dual inhibitors. There are two possible reasons for the misidentification of a substantial percentage of non-dual inhibitors as dual-inhibitors. First, SVR QSAR models were trained by non-dual inhibitors only, which may not fully distinguish dual and non-dual inhibitors. Secondly, some of the misidentified non-dual inhibitors are probably true dual-inhibitors not yet experimentally tested for multi-target activities. It is noted that "mistaken" selection of these non-dual inhibitors is still useful for searching single-target leads.

Target selectivity was further tested by using SVR QSAR models to screen the 1,816-4,298 inhibitors of the other 3 kinases not included in a particular kinase-pair. We found that 33.9% of these inhibitors were misidentified as dual-inhibitors for EGFR-VEGFR, 18.3% for EGFR-PDGFR, 7.1% for EGFR-FGFR and 12.5% for EGFR-Src respectively. These showed that our SVR QSAR models are fairly selective in separating inhibitors of specific kinase pair from those of other kinases.

Virtual-hit rates and false-hit rates of our SVR QSAR method in screening compounds that resemble the structural and physicochemical properties of the training datasets were evaluated by using 1,175-9,356 MDDR compounds similar to a dual-inhibitor of each kinase pair. Similarity was defined by Tanimoto similarity coefficient $\geq$0.9 between a MDDR compound and its closest dual-inhibitor.[33] Our SVR QSAR models identified 691 virtual-hits from 9,356 MDDR similarity compounds (virtual-hit rate 7.39%) for EGFR-VEGFR, 194 from 1,175 MDDR compounds (16.51%) for EGFR-PDGFR, 85 from 1,285 MDDR compounds (6.61%) for EGFR-FGFR, and 319 from 5,404 MDDR compounds (5.90%) for EGFR-Src respectively.

Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168 thousand MDDR and 13.56 million PubChem compounds. The numbers of virtual hits and virtual-hit rates in screening 168 thousand MDDR compounds are 2,605 and 1.55% for EGFR-VEGFR, 1,557 and 0.93% for EGFR-PDGFR, 654 and 0.39% for EGFR-FGFR, and 1,656 and 0.99% for EGFR-Src respectively. The numbers of virtual hits and virtual-hit rates in screening 13.56M PubChem compounds are 102,497 and 0.74% for EGFR-VEGFR, 61,764 and 0.46% for EGFR-PDGFR, 18,981 and 0.14% for EGFR-FGFR, and 52,498 and 0.39% for EGFR-Src respectively.

Substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, tyrosine-specific protein kinase inhibitors, and signal transduction inhibitors (**Table 5.3**). As some of these virtual-hits may be true dual-inhibitors, the false-hit rates of our SVR QSAR models are at most equal to and likely less than the virtual-hit rates. Hence the false-hit rates are satisfactorily low with ≤6.61%-16.51% in screening 1,175-9,356 MDDR similarity compounds, ≤0.39%-1.55% in screening 168 thousand MDDR compounds, and ≤0.14%-0.74% in screening 13.56 million PubChem compounds, which are comparable and in some cases better than single-target false-hit rates of 0.0054%-8.3% of single-target support vector machine (SVM) methods,[276, 296] 0.08%-3% of structure-based methods, 0.1%-5% by other machine learning methods, 0.16%-8.2% by clustering methods, and 1.15%-26% by pharmacophore models.[313]

5.3.3 Evaluation of SVR QSAR models identified MDDR virtual hits

Our SVR QSAR models identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 5.3** gives the MDDR classes that contain higher percentage (≥5%) of SVR QSAR virtual hits and the percentage values. We found that 248-1,092 or 36.4%-41.9% of the 654-2,605 virtual hits belong to the antineoplastic class, which represent 1.3%-5.6% of the 19,643 MDDR compounds in the class. In particular, 67-341 or 10.2%-14.8% of the virtual hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 5.7%-28.9% of the 1,181 MDDR compounds in the class. Moreover, 76-268 or 9.9%-13.8% of the virtual hits belong to the signal transduction inhibitor class, representing 3.7%-13.2% of the 2,037 members in this class. Therefore, many of the SVR QSAR virtual hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases involved in signal transduction, angiogenesis and other cancer-related pathways. While some of these kinase inhibitors might be true dual-inhibitors of specific kinase-pairs, the majority of them are expected to arise from false selection of non-dual inhibitors of the same kinase-pairs (at 13.6%-60.2% false-hit rates) and inhibitors of other kinases (at 7.1%-33.9% false-hit rates).

Some of the SVR QSAR virtual hits belong to the antiarthritic class. All of our evaluated kinases or their kinase-likes have been linked to arthritis in the literature. EGFR-like receptor stimulates synovial cells and its elevated activities may be involved in the pathogenesis of rheumatoid arthritis.[296] VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis.[314] FGFR may partly mediate osteoarthritis.[315] PDGF-like factors stimulates the proliferative and invasive phenotype of rheumatoid arthritis synovial connective tissue cells.[316] Therefore, some of the SVR QSAR virtual hits in the antiarthritic class may be inhibitors of our evaluated kinases or their kinase-likes capable of producing antiarthritic activities.

Table 5.3 MDDR classes that contain higher percentage (≥5%) of virtual-hits identified by combinatorial SVMs in screening 168 thousand MDDR compounds for dual-inhibitors of 4 combinations of EGFR, VEGFR, PDGFR, FGFR and Src.

Kinase Pair	No of SVR Identified Virtual Hits	MDDR Classes that Contain Higher Percentage of Virtual Hits	No of Virtual Hits in Class	Percentage of Class member as Virtual Hits
EGFR-VEGFR	2,605	Antineoplastic	1,092	41.9%
		Tyrosine-Specific Protein Kinase Inhibitor	341	13.1%
		Antiarthritic	298	11.4%
		Signal Transduction Inhibitor	268	10.3%
		Antiallergic/Antiasthmatic	148	5.7%
EGFR-PDGFR	1,557	Antineoplastic	566	36.4%
		Tyrosine-Specific Protein Kinase Inhibitor	209	13.4%
		Antiarthritic	180	11.6%
		Signal Transduction Inhibitor	154	9.9%
		Antiallergic/Antiasthmatic	107	6.9%
EGFR-FGFR	654	Antineoplastic	248	37.9%
		Antiarthritic	76	11.6%
		Signal Transduction Inhibitor	76	11.6%
		Tyrosine-Specific Protein Kinase Inhibitor	67	10.2%
		Antihypertensive	42	6.4%
EGFR-Src	1,656	Antineoplastic	677	40.9%
		Tyrosine-Specific Protein Kinase Inhibitor	245	14.8%
		Signal Transduction Inhibitor	228	13.8%
		Antiarthritic	174	10.5%
		Cephalosporin	112	6.8%

5.4 Further perspective

The high throughput SVR QSAR VS tools developed by using non-dual inhibitors show good capability in identifying dual-inhibitors of several anticancer target kinase-pairs at comparable and in many cases substantially lower false-hit rates than those of typical VS tools reported in the literature. The capability of the SVR QSAR models and other VS tools in identifying multi-kinase inhibitors and other multi-target agents may be further enhanced by incorporating knowledge of multi-target agents into VS tool development processes. With the discovery of increasing number of selective multi-target agents from the current and future drug discovery efforts, it is possible to introduce more comprehensive elements of distinguished structural and physicochemical features of selective multi-target agents into the training of combinatorial VS tools for more effective identification of selective multi-target agents. These multi-target VS tools may be combined with structure-based filters for enhanced target selectivity. Because of the high computing speed and generalization capability, our SVR QSAR method can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of multi-kinase inhibitors and other multi-target agents.

CHAPTER 6 Concluding Remarks

This last chapter summarizes the major findings and contributions of this study (**Section 6.1**). Limitations of present study and suggestions on possible areas for further studies are discussed in **Section 6.2**.

6.1 Major findings and contributions

In this work, a Pathway Cross-talk Database (PCD) was developed providing information on experimentally confirmed pathway cross-talks with detailed information about the interactive mediators and mechanisms. PCD currently contains 137 entries of experimentally discovered pathway cross-talks described in the literature. There are a total of 89 pathways or pathway components covering 78 diseases or biological processes included in the database. Rapid advances in the study of systems level regulations and cross-talks and in the investigation of their molecular mechanisms are expected to generate more information and stimulate more interest in exploring pathway cross-talks for regulating biological processes via chemical and other means, and for discovering multi-targeting drugs and drug combinations. By incorporating the relevant information generated from these studies, PCD may complement and expand the application scope of other pathway databases to facilitate systems-level studies of biological regulations and disease processes, and the discovery of multi-targeting drugs and drug combinations. At last, four combinations of five kinases, EGFR-VEGFR, EGFR-PDGFR, EGFR-FGFR and EGFR-Src, have been identified as promising targets for treating NSCLC.

Machine learning (ML) methods have been explored for developing QSAR models as alternative VS tools searching single- and multi-target agents because of their high-CPU speed and capability for covering highly diverse spectrum of compounds. However, while exhibiting equally good hit selection and activity assessment performance in screening large libraries, the currently developed ML QSAR VS tools cannot identify highly novel inhibitors outside similarity-based ADs. In this work, a high throughput QSAR approach was developed using support vector regression (SVR) as the regression algorithm and tested whether the performance of SVR QSAR models can be improved by using training sets of diverse inactive compounds. Apart from the use of known inactive compounds and active compounds of other biological target classes as putative inactive compounds, an in-house algorithm was applied for generating putative inactive compounds. An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the “inactive” chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. Our models performed well in predicting new inhibitors reported after the year of 2010 with R^2 values comparable to those of other QSAR models. In retrospective database screening of active compounds from large libraries such as PubChem and MDDR, our SVR QSAR models also showed improved hit-rates and the enrichment factors. Moreover, our method showed some level of capability in the identification and activity assessment of highly novel inhibitors outside similarity-based ADs (as summarized in **Table 6.1**). The putative negatives generation method plays an important role in it. This method greatly increased the performance of VS without compromising performance within ADs. It showed that at the study of chemistry and biological problems, certain assumption could be made to solve the problems although sometimes it may lead to certain degree of noise.

Our SVR QSAR models were tested as VS tools for searching dual-inhibitors of 4 combinations of 5 anticancer kinase targets (EGFR, VEGFR, PDGFR, FGFR and Src). SVR QSAR Models were fairly selective in misidentifying as dual-inhibitors of the non-dual inhibitors of the same kinase-pairs and produced low false-hit rates in misidentifying as dual-inhibitors of PubChem and MDDR databases. Compared with other methods, our SVR QSAR models show good capability in identifying dual-inhibitors of several anticancer target kinase-pairs at comparable and in many cases substantially lower false-hit rates. Therefore, SVR QSAR models are potentially useful to discover multi-target agents for enhancing efficacy and reducing counter-target activities and toxicities.

Table 6.1 Comparison of the SVR QSAR method with other established QSAR methods

QSAR methods	Regression method	Dataset	Application
Traditional QSAR	Equations	Only deal with small fraction of compounds (usually up to 100) similar to each other	Only applicable for lead optimization
Modern QSAR	Linear & non-linear ML (e.g. kNN, ANN, etc.)	Able to deal with larger number of compounds (with majority are active ones & very few inactive ones)	Only applicable for prediction on new compounds within similarity-based AD
SVR QSAR as in this work	SVR	Able to deal with large number of compounds (with extensive collection of active and inactive ones & putative negatives)	Applicable for prediction on new compounds within and beyond AD; able to scan large chemical database with satisfactory hit-rates and enrichment factors and low false-hit rates

6.2 Limitations and suggestions for future studies

The Pathway Cross-talk Database (PCD) is potentially useful for facilitating the systems level understanding of diseases, biological processes and treatment strategies. However, recently we realized that the old literature searching strategy was flawed during the database information collection step. In the year 2007 to 2008 when we first tried to develop this database, we used the keyword “crosstalk” combined with either “pathway” or “network” or “protein” to identify the literature that describe experimentally discovered cross-talk between two different pathways. However, the word “crosstalk” is only one way but not the most common. A PubMed search of “cross-talk” combined with “pathway”, for instance, results in over twice as many entries as “crosstalk” combined with “pathway”. This is one example that the old strategy was inadequate which resulted in a lot of relevant literature or data that should be collected in the database missed out and made this database an under-representation of the experimentally confirmed pathway cross-talks. Therefore, the current version of the database and the old strategy we used can only be seen as a prototype of a potential route towards a future comprehensive pathway cross-talk database. The searching strategy needs to be improved, for example, by adopting more proper keyword terms, aside from the old ones, such as "cross-talk" or "interaction" or "linkage" combined with "pathway" or "network".

On the other hand, it has been years since PCD was developed. It is now out of date because many useful papers have been published since then. For example, over 800 new papers were published since 2009 by searching PubMed using the term "crosstalk AND pathway". Thus new entries from the new papers in recent years will also be added to make this database up to date.

The SVR QSAR models developed using our putative negative dataset are not perfect. There are still some false hits that cannot be ruled out easily. These false hits are “correctly” identified by our SVR QSAR models due to the similar structural frameworks with real active compounds. Our molecular descriptors used in the SVR QSAR models are insufficient to adequately differentiate the compounds with similar structural frameworks. Therefore, it is necessary to explore different combinations of descriptors and to select any more optimal sets of descriptors by using more refined feature selection algorithms and parameters in future work. It may also be helpful to introduce new descriptors for more appropriate representation of compounds or descriptors which can be used to describe the interaction between targets and the ligands.

The putative negatives generation method helps a lot in improving the performance of SVR QSAR models in VS large chemical libraries. However, a drawback of this approach lies in the possible inclusion of some undiscovered active compounds in the "inactive" class, which may affect the capability of ML methods for identifying novel active compounds. As will be demonstrated, such an adverse effect is expected to be relatively small for many biological target classes. On the other hand, the clustering of chemical space can also affect the generation of putative negative dataset. Chemical space clustering is a difficult area in cheminformatics that is clustering method, distance metrix selection and molecular descriptors dependent. K-means clustering method used in this work is not the best clustering method but is suitable and computable for large chemical spaces. In future studies, new clustering algorithm can be developed for improving the accuracy of chemical space clustering. The selection of correlation coefficients and other chemical descriptors such as fingerprint also can be the direction of improvement.

Our SVR QSAR models showed the good performance in VS large chemical libraries with improved hit rate, yield and enrichment factor. Furthermore, our SVR QSAR models also showed some capability in identifying highly novel actives beyond similarity-based ADs. At this point, experimental studies are necessary for validating our high performance virtual screening tools. Based on this, we have formed extensive collaborations with several research groups and some compounds have been selected and sent to our collaborators for further study.

The capability of the SVR QSAR models in identifying multi-kinase inhibitors and other multi-target agents needs to be further enhanced by incorporating knowledge of multi-target agents into VS tool development processes. With the discovery of increasing number of selective multi-target agents from the current and future drug discovery efforts, it is possible to introduce more comprehensive elements of distinguished structural and physicochemical features of selective multi-target agents into the training of combinatorial VS tools for more effective identification of selective multi-target agents.

These years have seen plenty of debate aimed to define which of the many VS approaches the best is. However, this question remains not answered conclusively. Each approach has its own advantages and drawbacks, and the choice of one or the other depends on the particular research question faced by the medicinal chemist. In terms of performance, ligand based methods tend to present better enrichment factors and higher speed serving as a more efficient methodologies to remove non active compounds while target based method provides a more straightforward picture of interactions between the drug and molecular target and a better prediction in terms of novel structures. Now synergistic, rational and synthetic combinations of different approaches make a possible trend for future drug discovery. Combined VS approach tends to include less costly approaches, usually ligand based VS, at the first stage, while the most demanding methods, usually docking, for the last stage when the original large compound library has been reduced to a manageable size.

BIBLIOGRAPHY

1. Petrelli A and Giordano S. *From single- to multi-target drugs in cancer therapy: when aspecificity becomes an advantage.* Curr Med Chem, 2008. **15**(5): p. 422-32.
2. Pennell NA and Lynch TJ, Jr. *Combined inhibition of the VEGFR and EGFR signaling pathways in the treatment of NSCLC.* Oncologist, 2009. **14**(4): p. 399-411.
3. Sawyers C. *Targeted cancer therapy.* Nature, 2004. **432**(7015): p. 294-7.
4. Gossage L and Eisen T. *Targeting multiple kinase pathways: a change in paradigm.* Clin Cancer Res, 2010. **16**(7): p. 1973-8.
5. Logue JS and Morrison DK. *Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy.* Genes Dev, 2012. **26**(7): p. 641-50.
6. Tabernero J. *The role of VEGF and EGFR inhibition: implications for combining anti-VEGF and anti-EGFR agents.* Mol Cancer Res, 2007. **5**(3): p. 203-20.
7. Ciardiello F, Caputo R, Damiano V, Troiani T, Vitagliano D, Carlomagno F, Veneziani BM, Fontanini G, Bianco AR, and Tortora G. *Antitumor effects of ZD6474, a small molecule vascular endothelial growth factor receptor tyrosine kinase inhibitor, with additional activity against epidermal growth factor receptor tyrosine kinase.* Clin Cancer Res, 2003. **9**(4): p. 1546-56.
8. De Luca A, Carotenuto A, Rachiglio A, Gallo M, Maiello MR, Aldinucci D, Pinto A, and Normanno N. *The role of the EGFR signaling in tumor microenvironment.* J Cell Physiol, 2008. **214**(3): p. 559-67.
9. Petit AM, Rak J, Hung MC, Rockwell P, Goldstein N, Fendly B, and Kerbel RS. *Neutralizing antibodies against epidermal growth factor and ErbB-2/neu receptor tyrosine kinases down-regulate vascular endothelial growth factor production by tumor cells in vitro and in vivo: angiogenic implications for signal transduction therapy of solid tumors.* Am J Pathol, 1997. **151**(6): p. 1523-30.
10. Viloria-Petit AM and Kerbel RS. *Acquired resistance to EGFR inhibitors: mechanisms and prevention strategies.* Int J Radiat Oncol Biol Phys, 2004. **58**(3): p. 914-26.
11. Kari C, Chan TO, Rocha de Quadros M, and Rodeck U. *Targeting the epidermal growth factor receptor in cancer: apoptosis takes center stage.* Cancer Res, 2003. **63**(1): p. 1-5.
12. Yakes FM, Chinratanalab W, Ritter CA, King W, Seelig S, and Arteaga CL. *Herceptin-induced inhibition of phosphatidylinositol-3 kinase and Akt Is required for antibody-mediated effects on p27, cyclin D1, and antitumor action.* Cancer Res, 2002. **62**(14): p. 4132-41.
13. Busse D, Yakes FM, Lenferink AE, and Arteaga CL. *Tyrosine kinase inhibitors: rationale, mechanisms of action, and implications for drug resistance.* Semin Oncol, 2001. **28**(5 Suppl 16): p. 47-55.
14. Shintani M, Okazaki A, Masuda T, Kawada M, Ishizuka M, Doki Y, Weinstein IB, and Imoto M. *Overexpression of cyclin DI contributes to malignant properties of esophageal tumor cells by increasing VEGF production and decreasing Fas expression.* Anticancer Res, 2002. **22**(2A): p. 639-47.
15. Apsel B, Blair JA, Gonzalez B, Nazif TM, Feldman ME, Aizenstein B, Hoffman R, Williams RL, Shokat KM, and Knight ZA. *Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases.* Nat Chem Biol, 2008. **4**(11): p. 691-9.
16. Keith CT, Borisy AA, and Stockwell BR. *Multicomponent therapeutics for networked systems.* Nat Rev Drug Discov, 2005. **4**(1): p. 71-8.

17. Smalley KS, Haass NK, Brafford PA, Lioni M, Flaherty KT, and Herlyn M. *Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases.* Mol Cancer Ther, 2006. **5**(5): p. 1136-44.

18. Larder BA, Kemp SD, and Harrigan PR. *Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy.* Science, 1995. **269**(5224): p. 696-9.

19. Zhang X, Crespo A, and Fernandez A. *Turning promiscuous kinase inhibitors into safer drugs.* Trends Biotechnol, 2008. **26**(6): p. 295-301.

20. Ma XH, Shi Z, Tan C, Jiang Y, Go ML, Low BC, and Chen YZ. *In-silico approaches to multi-target drug discovery : computer aided multi-target drug design, multi-target virtual screening.* Pharm Res, 2010. **27**(5): p. 739-49.

21. Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, Zhang L, Song Y, Zhang J, Han B, Zhang P, and Chen Y. *Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1128-36.

22. Tse C, Shoemaker AR, Adickes J, Anderson MG, Chen J, Jin S, Johnson EF, Marsh KC, Mitten MJ, Nimmer P, Roberts L, Tahir SK, Xiao Y, Yang X, Zhang H, Fesik S, Rosenberg SH, and Elmore SW. *ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor.* Cancer Res, 2008. **68**(9): p. 3421-8.

23. Lock R, Carol H, Houghton PJ, Morton CL, Kolb EA, Gorlick R, Reynolds CP, Maris JM, Keir ST, Wu J, and Smith MA. *Initial testing (stage 1) of the BH3 mimetic ABT-263 by the pediatric preclinical testing program.* Pediatr Blood Cancer, 2008. **50**(6): p. 1181-9.

24. Reid A, Vidal L, Shaw H, and de Bono J. *Dual inhibition of ErbB1 (EGFR/HER1) and ErbB2 (HER2/neu).* Eur J Cancer, 2007. **43**(3): p. 481-9.

25. Takezawa K, Okamoto I, Tanizaki J, Kuwata K, Yamaguchi H, Fukuoka M, Nishio K, and Nakagawa K. *Enhanced anticancer effect of the combination of BIBW2992 and thymidylate synthase-targeted agents in non-small cell lung cancer with the T790M mutation of epidermal growth factor receptor.* Mol Cancer Ther, 2010. **9**(6): p. 1647-56.

26. Howard S, Berdini V, Boulstridge JA, Carr MG, Cross DM, Curry J, Devine LA, Early TR, Fazal L, Gill AL, Heathcote M, Maman S, Matthews JE, McMenamin RL, Navarro EF, O'Brien MA, O'Reilly M, Rees DC, Reule M, Tisi D, Williams G, Vinkovic M, and Wyatt PG. *Fragment-based discovery of the pyrazol-4-yl urea (AT9283), a multitargeted kinase inhibitor with potent aurora kinase activity.* J Med Chem, 2009. **52**(2): p. 379-88.

27. Curry J, Angove H, Fazal L, Lyons J, Reule M, Thompson N, and Wallis N. *Aurora B kinase inhibition in mitosis: strategies for optimising the use of aurora kinase inhibitors such as AT9283.* Cell Cycle, 2009. **8**(12): p. 1921-9.

28. Gill AL, Verdonk M, Boyle RG, and Taylor R. *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): p. 1408-22.

29. Hu-Lowe DD, Zou HY, Grazzini ML, Hallin ME, Wickman GR, Amundson K, Chen JH, Rewolinski DA, Yamazaki S, Wu EY, McTigue MA, Murray BW, Kania RS, O'Connor P, Shalinsky DR, and Bender SL. *Nonclinical antiangiogenesis and antitumor activities of axitinib (AG-013736), an oral, potent, and selective inhibitor of vascular endothelial growth factor receptor tyrosine kinases 1, 2, 3.* Clin Cancer Res, 2008. **14**(22): p. 7272-83.

30. Rossler J, Monnet Y, Farace F, Opolon P, Daudigeos-Dubus E, Bourredjem A, Vassal G, and Geoerger B. *The selective VEGFR1-3 inhibitor axitinib (AG-013736) shows antitumor activity in human neuroblastoma xenografts.* Int J Cancer, 2011. **128**(11): p. 2748-58.

31. Hennequin LF, Allen J, Breed J, Curwen J, Fennell M, Green TP, Lambert-van der Brempt C, Morgentin R, Norman RA, Olivier A, Otterbein L, Ple PA, Warin N, and Costello G. *N-(5-chloro-*

1,3-benzodioxol-4-yl)-7-[2-(4-methylpiperazin-1-yl)ethoxy]-5- (tetrahydro-2H-pyran-4-yloxy)quinazolin-4-amine, a novel, highly selective, orally available, dual-specific c-Src/Abl kinase inhibitor. J Med Chem, 2006. **49**(22): p. 6465-88.

32. Arcaroli JJ, Touban BM, Tan AC, Varella-Garcia M, Powell RW, Eckhardt SG, Elvin P, Gao D, and Messersmith WA. *Gene array and fluorescence in situ hybridization biomarkers of activity of saracatinib (AZD0530), a Src inhibitor, in a preclinical model of colorectal cancer.* Clin Cancer Res, 2010. **16**(16): p. 4165-77.

33. Koppikar P, Choi SH, Egloff AM, Cai Q, Suzuki S, Freilino M, Nozawa H, Thomas SM, Gooding WE, Siegfried JM, and Grandis JR. *Combined inhibition of c-Src and epidermal growth factor receptor abrogates growth and invasion of head and neck squamous cell carcinoma.* Clin Cancer Res, 2008. **14**(13): p. 4284-91.

34. Wojtowicz-Praga SM, Dickson RB, and Hawkins MJ. *Matrix metalloproteinase inhibitors.* Invest New Drugs, 1997. **15**(1): p. 61-75.

35. Low JA, Johnson MD, Bone EA, and Dickson RB. *The matrix metalloproteinase inhibitor batimastat (BB-94) retards human breast cancer solid tumor growth but not ascites formation in nude mice.* Clin Cancer Res, 1996. **2**(7): p. 1207-14.

36. Wong TW, Lee FY, Yu C, Luo FR, Oppenheimer S, Zhang H, Smykla RA, Mastalerz H, Fink BE, Hunt JT, Gavai AV, and Vite GD. *Preclinical antitumor activity of BMS-599626, a pan-HER kinase inhibitor that inhibits HER1/HER2 homodimer and heterodimer signaling.* Clin Cancer Res, 2006. **12**(20 Pt 1): p. 6186-93.

37. Golas JM, Arndt K, Etienne C, Lucas J, Nardin D, Gibbons J, Frost P, Ye F, Boschelli DH, and Boschelli F. *SKI-606, a 4-anilino-3-quinolinecarbonitrile dual inhibitor of Src and Abl kinases, is a potent antiproliferative agent against chronic myelogenous leukemia cells in culture and causes regression of K562 xenografts in nude mice.* Cancer Res, 2003. **63**(2): p. 375-81.

38. Zhang YX, Knyazev PG, Cheburkin YV, Sharma K, Knyazev YP, Orfi L, Szabadkai I, Daub H, Keri G, and Ullrich A. *AXL is a potential target for therapeutic intervention in breast cancer progression.* Cancer Res, 2008. **68**(6): p. 1905-15.

39. Carroll FI, Muresan AZ, Blough BE, Navarro HA, Mascarella SW, Eaton JB, Huang X, Damaj MI, and Lukas RJ. *Synthesis of 2-(substituted phenyl)-3,5,5-trimethylmorpholine analogues and their effects on monoamine uptake, nicotinic acetylcholine receptor function, and behavioral effects of nicotine.* J Med Chem, 2011. **54**(5): p. 1441-8.

40. Ferris RM and Beaman OJ. *Bupropion: a new antidepressant drug, the mechanism of action of which is not associated with down-regulation of postsynaptic beta-adrenergic, serotonergic (5-HT2), alpha 2-adrenergic, imipramine and dopaminergic receptors in brain.* Neuropharmacology, 1983. **22**(11): p. 1257-67.

41. Fryer JD and Lukas RJ. *Noncompetitive functional inhibition at diverse, human nicotinic acetylcholine receptor subtypes by bupropion, phencyclidine, and ibogaine.* J Pharmacol Exp Ther, 1999. **288**(1): p. 88-92.

42. Rabindran SK, Discafani CM, Rosfjord EC, Baxter M, Floyd MB, Golas J, Hallett WA, Johnson BD, Nilakantan R, Overbeek E, Reich MF, Shen R, Shi X, Tsou HR, Wang YF, and Wissner A. *Antitumor activity of HKI-272, an orally active, irreversible inhibitor of the HER-2 tyrosine kinase.* Cancer Res, 2004. **64**(11): p. 3958-65.

43. Buchdunger E, Zimmermann J, Mett H, Meyer T, Muller M, Druker BJ, and Lydon NB. *Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative.* Cancer Res, 1996. **56**(1): p. 100-4.

44. Heinrich MC, Griffith DJ, Druker BJ, Wait CL, Ott KA, and Zigler AJ. *Inhibition of c-kit receptor tyrosine kinase activity by STI 571, a selective tyrosine kinase inhibitor.* Blood, 2000. **96**(3): p. 925-32.

45. Radujkovic A, Schad M, Topaly J, Veldwijk MR, Laufs S, Schultheis BS, Jauch A, Melo JV, Fruehauf S, and Zeller WJ. *Synergistic activity of imatinib and 17-AAG in imatinib-resistant CML cells overexpressing BCR-ABL--Inhibition of P-glycoprotein function by 17-AAG.* Leukemia, 2005. **19**(7): p. 1198-206.

46. Rusnak DW, Lackey K, Affleck K, Wood ER, Alligood KJ, Rhodes N, Keith BR, Murray DM, Knight WB, Mullin RJ, and Gilmer TM. *The effects of the novel, reversible epidermal growth factor receptor/ErbB-2 tyrosine kinase inhibitor, GW2016, on the growth of human normal and tumor-derived cell lines in vitro and in vivo.* Mol Cancer Ther, 2001. **1**(2): p. 85-94.

47. Sawyers CL. *Finding the next Gleevec: FLT3 targeted kinase inhibitor therapy for acute myeloid leukemia.* Cancer Cell, 2002. **1**(5): p. 413-5.

48. Budworth J, Davies R, Malkhandi J, Gant TW, Ferry DR, and Gescher A. *Comparison of staurosporine and four analogues: their effects on growth, rhodamine 123 retention and binding to P-glycoprotein in multidrug-resistant MCF-7/Adr cells.* Br J Cancer, 1996. **73**(9): p. 1063-8.

49. Gleixner KV, Rebuzzi L, Mayerhofer M, Gruze A, Hadzijusufovic E, Sonneck K, Vales A, Kneidinger M, Samorapoompichit P, Thaiwong T, Pickl WF, Yuzbasiyan-Gurkan V, Sillaber C, Willmann M, and Valent P. *Synergistic antiproliferative effects of KIT tyrosine kinase inhibitors on neoplastic canine mast cells.* Exp Hematol, 2007. **35**(10): p. 1510-21.

50. Huang YC, Chao DK, Clifford Chao KS, and Chen YJ. *Oral small-molecule tyrosine kinase inhibitor midostaurin (PKC412) inhibits growth and induces megakaryocytic differentiation in human leukemia cells.* Toxicol In Vitro, 2009. **23**(6): p. 979-85.

51. Shimomura T, Hasako S, Nakatsuru Y, Mita T, Ichikawa K, Kodera T, Sakai T, Nambu T, Miyamoto M, Takahashi I, Miki S, Kawanishi N, Ohkubo M, Kotani H, and Iwasawa Y. *MK-5108, a highly selective Aurora-A kinase inhibitor, shows antitumor activity alone and in combination with docetaxel.* Mol Cancer Ther, 2010. **9**(1): p. 157-66.

52. Polverino A, Coxon A, Starnes C, Diaz Z, DeMelfi T, Wang L, Bready J, Estrada J, Cattley R, Kaufman S, Chen D, Gan Y, Kumar G, Meyer J, Neervannan S, Alva G, Talvenheimo J, Montestruque S, Tasker A, Patel V, Radinsky R, and Kendall R. *AMG 706, an oral, multikinase inhibitor that selectively targets vascular endothelial growth factor, platelet-derived growth factor, and kit receptors, potently inhibits angiogenesis and induces regression in tumor xenografts.* Cancer Res, 2006. **66**(17): p. 8715-21.

53. Coxon A, Bush T, Saffran D, Kaufman S, Belmontes B, Rex K, Hughes P, Caenepeel S, Rottman JB, Tasker A, Patel V, Kendall R, Radinsky R, and Polverino A. *Broad antitumor activity in breast cancer xenografts by motesanib, a highly selective, oral inhibitor of vascular endothelial growth factor, platelet-derived growth factor, and Kit receptors.* Clin Cancer Res, 2009. **15**(1): p. 110-8.

54. Weisberg E, Manley PW, Breitenstein W, Bruggen J, Cowan-Jacob SW, Ray A, Huntly B, Fabbro D, Fendrich G, Hall-Meyers E, Kung AL, Mestan J, Daley GQ, Callahan L, Catley L, Cavazza C, Azam M, Neuberg D, Wright RD, Gilliland DG, and Griffin JD. *Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl.* Cancer Cell, 2005. **7**(2): p. 129-41.

55. Petti F, Thelemann A, Kahler J, McCormack S, Castaldo L, Hunt T, Nuwaysir L, Zeiske L, Haack H, Sullivan L, Garton A, and Haley JD. *Temporal quantitation of mutant Kit tyrosine kinase signaling attenuated by a novel thiophene kinase inhibitor OSI-930.* Mol Cancer Ther, 2005. **4**(8): p. 1186-97.

56. Garton AJ, Crew AP, Franklin M, Cooke AR, Wynne GM, Castaldo L, Kahler J, Winski SL, Franks A, Brown EN, Bittner MA, Keily JF, Briner P, Hidden C, Srebernak MC, Pirrit C, O'Connor M, Chan A, Vulevic B, Henninger D, Hart K, Sennello R, Li AH, Zhang T, Richardson F, Emerson DL, Castelhano AL, Arnold LD, and Gibson NW. *OSI-930: a novel selective inhibitor of Kit and kinase insert domain receptor tyrosine kinases with antitumor activity in mouse xenograft models.* Cancer Res, 2006. **66**(2): p. 1015-24.

57. Joshi KS, Rathos MJ, Joshi RD, Sivakumar M, Mascarenhas M, Kamble S, Lal B, and Sharma S. *In vitro antitumor properties of a novel cyclin-dependent kinase inhibitor, P276-00.* Mol Cancer Ther, 2007. **6**(3): p. 918-25.

58. Manohar SM, Rathos MJ, Sonawane V, Rao SV, and Joshi KS. *Cyclin-dependent kinase inhibitor, P276-00 induces apoptosis in multiple myeloma cells by inhibition of Cdk9-T1 and RNA polymerase II-dependent transcription.* Leuk Res, 2011. **35**(6): p. 821-30.

59. Bruns C, Lewis I, Briner U, Meno-Tetang G, and Weckbecker G. *SOM230: a novel somatostatin peptidomimetic with broad somatotropin release inhibiting factor (SRIF) receptor binding and a unique antisecretory profile.* Eur J Endocrinol, 2002. **146**(5): p. 707-16.

60. Adams RL, Adams IP, Lindow SW, and Atkin SL. *Inhibition of endothelial proliferation by the somatostatin analogue SOM230.* Clin Endocrinol (Oxf), 2004. **61**(4): p. 431-6.

61. Ivy SP, Wick JY, and Kaufman BM. *An overview of small-molecule inhibitors of VEGFR signaling.* Nat Rev Clin Oncol, 2009. **6**(10): p. 569-79.

62. Kumar R, Knick VB, Rudolph SK, Johnson JH, Crosby RM, Crouthamel MC, Hopper TM, Miller CG, Harrington LE, Onori JA, Mullin RJ, Gilmer TM, Truesdale AT, Epperly AH, Boloor A, Stafford JA, Luttrell DK, and Cheung M. *Pharmacokinetic-pharmacodynamic correlation from mouse to human with pazopanib, a multikinase angiogenesis inhibitor with potent antitumor and antiangiogenic activity.* Mol Cancer Ther, 2007. **6**(7): p. 2012-21.

63. Jani JP, Arcari J, Bernardo V, Bhattacharya SK, Briere D, Cohen BD, Coleman K, Christensen JG, Emerson EO, Jakowski A, Hook K, Los G, Moyer JD, Pruimboom-Brees I, Pustilnik L, Rossi AM, Steyn SJ, Su C, Tsaparikos K, Wishka D, Yoon K, and Jakubczak JL. *PF-03814735, an orally bioavailable small molecule aurora kinase inhibitor for cancer therapy.* Mol Cancer Ther, 2010. **9**(4): p. 883-94.

64. Carpinelli P, Ceruti R, Giorgini ML, Cappella P, Gianellini L, Croci V, Degrassi A, Texido G, Rocchetti M, Vianello P, Rusconi L, Storici P, Zugnoni P, Arrigoni C, Soncini C, Alli C, Patton V, Marsiglio A, Ballinari D, Pesenti E, Fancelli D, and Moll J. *PHA-739358, a potent inhibitor of Aurora kinases with a selective target inhibition profile relevant to cancer.* Mol Cancer Ther, 2007. **6**(12 Pt 1): p. 3158-68.

65. Chen R, Wierda WG, Chubb S, Hawtin RE, Fox JA, Keating MJ, Gandhi V, and Plunkett W. *Mechanism of action of SNS-032, a novel cyclin-dependent kinase inhibitor, in chronic lymphocytic leukemia.* Blood, 2009. **113**(19): p. 4637-45.

66. Conroy A, Stockett DE, Walker D, Arkin MR, Hoch U, Fox JA, and Hawtin RE. *SNS-032 is a potent and selective CDK 2, 7 and 9 inhibitor that drives target modulation in patient samples.* Cancer Chemother Pharmacol, 2009. **64**(4): p. 723-32.

67. Wilhelm SM, Carter C, Tang L, Wilkie D, McNabola A, Rong H, Chen C, Zhang X, Vincent P, McHugh M, Cao Y, Shujath J, Gawlak S, Eveleigh D, Rowley B, Liu L, Adnane L, Lynch M, Auclair D, Taylor I, Gedrich R, Voznesensky A, Riedl B, Post LE, Bollag G, and Trail PA. *BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis.* Cancer Res, 2004. **64**(19): p. 7099-109.

68. Plaza-Menacho I, Mologni L, Sala E, Gambacorti-Passerini C, Magee AI, Links TP, Hofstra RM, Barford D, and Isacke CM. *Sorafenib functions to potently suppress RET tyrosine kinase activity by direct enzymatic inhibition and promoting RET lysosomal degradation independent of proteasomal targeting.* J Biol Chem, 2007. **282**(40): p. 29230-40.

69. Liu L, Cao Y, Chen C, Zhang X, McNabola A, Wilkie D, Wilhelm S, Lynch M, and Carter C. *Sorafenib blocks the RAF/MEK/ERK pathway, inhibits tumor angiogenesis, and induces tumor cell apoptosis in hepatocellular carcinoma model PLC/PRF/5.* Cancer Res, 2006. **66**(24): p. 11851-8.

70. Auclair D, Miller D, Yatsula V, Pickett W, Carter C, Chang Y, Zhang X, Wilkie D, Burd A, Shi H, Rocks S, Gedrich R, Abriola L, Vasavada H, Lynch M, Dumas J, Trail PA, and Wilhelm SM. *Antitumor activity of sorafenib in FLT3-driven leukemic cells.* Leukemia, 2007. **21**(3): p. 439-45.

71. Skvara H, Dawid M, Kleyn E, Wolff B, Meingassner JG, Knight H, Dumortier T, Kopp T, Fallahi N, Stary G, Burkhart C, Grenet O, Wagner J, Hijazi Y, Morris RE, McGeown C, Rordorf C, Griffiths CE, Stingl G, and Jung T. *The PKC inhibitor AEB071 may be a therapeutic option for psoriasis.* J Clin Invest, 2008. **118**(9): p. 3151-9.

72. Evenou JP, Wagner J, Zenke G, Brinkmann V, Wagner K, Kovarik J, Welzenbach KA, Weitz-Schmidt G, Guntermann C, Towbin H, Cottens S, Kaminski S, Letschka T, Lutz-Nicoladoni C, Gruber T, Hermann-Kleiter N, Thuille N, and Baier G. *The potent protein kinase C-selective inhibitor AEB071 (sotrastaurin) represents a new class of immunosuppressive agents affecting early T-cell activation.* J Pharmacol Exp Ther, 2009. **330**(3): p. 792-801.

73. Godl K, Gruss OJ, Eickhoff J, Wissing J, Blencke S, Weber M, Degen H, Brehmer D, Orfi L, Horvath Z, Keri G, Muller S, Cotten M, Ullrich A, and Daub H. *Proteomic characterization of the angiogenesis inhibitor SU6668 reveals multiple impacts on cellular kinase signaling.* Cancer Res, 2005. **65**(15): p. 6919-26.

74. Laird AD, Vajkoczy P, Shawver LK, Thurnher A, Liang C, Mohammadi M, Schlessinger J, Ullrich A, Hubbard SR, Blake RA, Fong TA, Strawn LM, Sun L, Tang C, Hawtin R, Tang F, Shenoy N, Hirth KP, McMahon G, and Cherrington. *SU6668 is a potent antiangiogenic and antitumor agent that induces regression of established tumors.* Cancer Res, 2000. **60**(15): p. 4152-60.

75. Krystal GW, Honsawek S, Kiewlich D, Liang C, Vasile S, Sun L, McMahon G, and Lipson KE. *Indolinone tyrosine kinase inhibitors block Kit activation and growth of small cell lung cancer cells.* Cancer Res, 2001. **61**(9): p. 3660-8.

76. Smolich BD, Yuen HA, West KA, Giles FJ, Albitar M, and Cherrington JM. *The antiangiogenic protein kinase inhibitors SU5416 and SU6668 inhibit the SCF receptor (c-kit) in a human myeloid leukemia cell line and in acute myeloid leukemia blasts.* Blood, 2001. **97**(5): p. 1413-21.

77. O'Farrell AM, Abrams TJ, Yuen HA, Ngai TJ, Louie SG, Yee KW, Wong LM, Hong W, Lee LB, Town A, Smolich BD, Manning WC, Murray LJ, Heinrich MC, and Cherrington JM. *SU11248 is a novel FLT3 tyrosine kinase inhibitor with potent activity in vitro and in vivo.* Blood, 2003. **101**(9): p. 3597-605.

78. Abrams TJ, Lee LB, Murray LJ, Pryer NK, and Cherrington JM. *SU11248 inhibits KIT and platelet-derived growth factor receptor beta in preclinical models of human small cell lung cancer.* Mol Cancer Ther, 2003. **2**(5): p. 471-8.

79. Sun L, Liang C, Shirazian S, Zhou Y, Miller T, Cui J, Fukuda JY, Chu JY, Nematalla A, Wang X, Chen H, Sistla A, Luu TC, Tang F, Wei J, and Tang C. *Discovery of 5-[5-fluoro-2-oxo-1,2-dihydroindol-(3Z)-ylidenemethyl]-2,4- dimethyl-1H-pyrrole-3-carboxylic acid (2-diethylaminoethyl)amide, a novel tyrosine kinase inhibitor targeting vascular endothelial and platelet-derived growth factor receptor tyrosine kinase.* J Med Chem, 2003. **46**(7): p. 1116-9.

80. Gozalbes R, Simon L, Froloff N, Sartori E, Monteils C, and Baudelle R. *Development and experimental validation of a docking strategy for the generation of kinase-targeted libraries.* J Med Chem, 2008. **51**(11): p. 3124-32.

81. Nagasawa J, Mizokami A, Koshida K, Yoshida S, Naito K, and Namiki M. *Novel HER2 selective tyrosine kinase inhibitor, TAK-165, inhibits bladder, kidney and androgen-independent prostate cancer in vitro and in vivo.* Int J Urol, 2006. **13**(5): p. 587-92.

82. Trudel S, Li ZH, Wei E, Wiesmann M, Chang H, Chen C, Reece D, Heise C, and Stewart AK. *CHIR-258, a novel, multitargeted tyrosine kinase inhibitor for the potential treatment of t(4;14) multiple myeloma.* Blood, 2005. **105**(7): p. 2941-8.

83. Harrington EA, Bebbington D, Moore J, Rasmussen RK, Ajose-Adeogun AO, Nakayama T, Graham JA, Demur C, Hercend T, Diu-Hercend A, Su M, Golec JM, and Miller KM. *VX-680, a potent and selective small-molecule inhibitor of the Aurora kinases, suppresses tumor growth in vivo.* Nat Med, 2004. **10**(3): p. 262-7.

84. Qian F, Engst S, Yamaguchi K, Yu P, Won KA, Mock L, Lou T, Tan J, Li C, Tam D, Lougheed J, Yakes FM, Bentzien F, Xu W, Zaks T, Wooster R, Greshock J, and Joly AH. *Inhibition of tumor cell growth, invasion, and metastasis by EXEL-2880 (XL880, GSK1363089), a novel inhibitor of HGF and VEGF receptor tyrosine kinases.* Cancer Res, 2009. **69**(20): p. 8009-16.

85. Siemeister G, Luecking U, Wagner C, Detjen K, Mc Coy C, and Bosslet K. *Molecular and pharmacodynamic characteristics of the novel multi-target tumor growth inhibitor ZK 304709.* Biomed Pharmacother, 2006. **60**(6): p. 269-72.

86. Scholz A, Wagner K, Welzel M, Remlinger F, Wiedenmann B, Siemeister G, Rosewicz S, and Detjen KM. *The oral multitarget tumour growth inhibitor, ZK 304709, inhibits growth of pancreatic neuroendocrine tumours in an orthotopic mouse model.* Gut, 2009. **58**(2): p. 261-70.

87. Ma XH, Wang R, Tan CY, Jiang YY, Lu T, Rao HB, Li XY, Go ML, Low BC, and Chen YZ. *Virtual Screening of Selective Multitarget Kinase Inhibitors by Combinatorial Support Vector Machines.* Mol Pharm, 2010. **7**(5): p. 1545-1560.

88. Shi Z, Ma XH, Qin C, Jia J, Jiang YY, Tan CY, and Chen YZ. *Combinatorial support vector machines approach for virtual screening of selective multi-target serotonin reuptake inhibitors from large compound libraries.* J Mol Graph Model, 2012. **32**: p. 49-66.

89. Hu Y and Bajorath J. *Molecular scaffolds with high propensity to form multi-target activity cliffs.* J Chem Inf Model, 2010. **50**(4): p. 500-10.

90. Zambrowicz BP and Sands AT. *Knockouts model the 100 best-selling drugs--will they model the next 100?* Nat Rev Drug Discov, 2003. **2**(1): p. 38-51.

91. Ohlstein EH, Ruffolo RR, Jr., and Elliott JD. *Drug discovery in the next millennium.* Annu Rev Pharmacol Toxicol, 2000. **40**: p. 177-91.

92. Hewitt DJ, Hargreaves RJ, Curtis SP, and Michelson D. *Challenges in analgesic drug development.* Clin Pharmacol Ther, 2009. **86**(4): p. 447-50.

93. Zheng C, Han L, Yap CW, Xie B, and Chen Y. *Progress and problems in the exploration of therapeutic targets.* Drug Discov Today, 2006. **11**(9-10): p. 412-20.

94. Rask-Andersen M, Almen MS, and Schioth HB. *Trends in the exploitation of novel drug targets.* Nat Rev Drug Discov, 2011. **10**(8): p. 579-90.

95. Aggarwal BB, Sethi G, Baladandayuthapani V, Krishnan S, and Shishodia S. *Targeting cell signaling pathways for drug discovery: an old lock needs a new key.* J Cell Biochem, 2007. **102**(3): p. 580-92.

96. Sebolt-Leopold JS and English JM. *Mechanisms of drug inhibition of signalling molecules.* Nature, 2006. **441**(7092): p. 457-62.

97. Ibrahim YH. *Regulation of IGF and integrin pathway crosstalk in breast cancer cells*, 2007, University of Minnesota.

98. Ibrahim YH and Yee D. *Insulin-like growth factor-I and breast cancer therapy.* Clin Cancer Res, 2005. **11**(2 Pt 2): p. 944s-50s.

99. Sachdev D and Yee D. *The IGF system and breast cancer.* Endocr Relat Cancer, 2001. **8**(3): p. 197-209.

100. Lambert AW, Ozturk S, and Thiagalingam S. *Integrin signaling in mammary epithelial cells and breast cancer.* ISRN Oncol, 2012. **2012**: p. 493283.

101. Soung YH, Clifford JL, and Chung J. *Crosstalk between integrin and receptor tyrosine kinase signaling in breast carcinoma progression.* BMB Rep, 2010. **43**(5): p. 311-8.

102. Baron V, Calleja V, Ferrari P, Alengrin F, and Van Obberghen E. *p125Fak focal adhesion kinase is a substrate for the insulin and insulin-like growth factor-I tyrosine kinase receptors.* J Biol Chem, 1998. **273**(12): p. 7162-8.

103. Lebrun P, Baron V, Hauck CR, Schlaepfer DD, and Van Obberghen E. *Cell adhesion and focal adhesion kinase regulate insulin receptor substrate-1 expression.* J Biol Chem, 2000. **275**(49): p. 38371-7.

104. Rizzo S, Bisi A, Bartolini M, Mancini F, Belluti F, Gobbi S, Andrisano V, and Rampa A. *Multi-target strategy to address Alzheimer's disease: design, synthesis and biological evaluation of new tacrine-based dimers.* Eur J Med Chem, 2011. **46**(9): p. 4336-43.

105. Mueller CA, Weinmann W, Dresen S, Schreiber A, and Gergov M. *Development of a multi-target screening analysis for 301 drugs using a QTrap liquid chromatography/tandem mass spectrometry system and automated library searching.* Rapid Commun Mass Spectrom, 2005. **19**(10): p. 1332-8.

106. Morphy R and Rankovic Z. *The physicochemical challenges of designing multiple ligands.* J Med Chem, 2006. **49**(16): p. 4961-70.

107. Chen YZ and Zhi DG. *Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule.* Proteins, 2001. **43**(2): p. 217-26.

108. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, Wang X, and Jiang H. *TarFisDock: a web server for identifying drug targets with docking approach.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W219-24.

109. Luo H, Chen J, Shi L, Mikailov M, Zhu H, Wang K, He L, and Yang L. *DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W492-8.

110. Kinnings SL, Xie L, Fung KH, Jackson RM, and Bourne PE. *The Mycobacterium tuberculosis drugome and its polypharmacological implications.* PLoS Comput Biol, 2010. **6**(11): p. e1000976.

111. Liu X, Ouyang S, Yu B, Liu Y, Huang K, Gong J, Zheng S, Li Z, Li H, and Jiang H. *PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W609-14.

112. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, and Shoichet BK. *Relating protein pharmacology by ligand chemistry.* Nat Biotechnol, 2007. **25**(2): p. 197-206.

113. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, and Roth BL. *Predicting new molecular targets for known drugs.* Nature, 2009. **462**(7270): p. 175-81.

114. Campillos M, Kuhn M, Gavin AC, Jensen LJ, and Bork P. *Drug target identification using side-effect similarity.* Science, 2008. **321**(5886): p. 263-6.

115. Marzaro G, Chilin A, Guiotto A, Uriarte E, Brun P, Castagliuolo I, Tonus F, and Gonzalez-Diaz H. *Using the TOPS-MODE approach to fit multi-target QSAR models for tyrosine kinases inhibitors.* Eur J Med Chem, 2011. **46**(6): p. 2185-92.

116. Garcia I, Fall Y, and Gomez G. *Using topological indices to predict anti-Alzheimer and anti-parasitic GSK-3 inhibitors by multi-target QSAR in silico screening.* Molecules, 2010. **15**(8): p. 5408-22.

117. Garcia I, Fall Y, Gomez G, and Gonzalez-Diaz H. *First computational chemistry multi-target model for anti-Alzheimer, anti-parasitic, anti-fungi, and anti-bacterial activity of GSK-3 inhibitors in vitro, in vivo, and in different cellular lines.* Mol Divers, 2011. **15**(2): p. 561-7.

118. Liu Q, Zhou H, Liu L, Chen X, Zhu R, and Cao Z. *Multi-target QSAR modelling in the analysis and design of HIV-HCV co-inhibitors: an in-silico study.* BMC Bioinformatics, 2011. **12**: p. 294.

119. Prado-Prado FJ, Uriarte E, Borges F, and Gonzalez-Diaz H. *Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs.* Eur J Med Chem, 2009. **44**(11): p. 4516-21.

120. Gonzalez-Diaz H and Prado-Prado FJ. *Unified QSAR and network-based computational chemistry approach to antimicrobials, part 1: multispecies activity models for antifungals.* J Comput Chem, 2008. **29**(4): p. 656-67.

121. Gonzalez-Diaz H, Prado-Prado FJ, Santana L, and Uriarte E. *Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species.* Bioorg Med Chem, 2006. **14**(17): p. 5973-80.

122. Morphy R. *The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds.* J Med Chem, 2006. **49**(10): p. 2969-78.

123. Jia J, Zhu F, Ma X, Cao Z, Li Y, and Chen YZ. *Mechanisms of drug combinations: interaction and network perspectives.* Nat Rev Drug Discov, 2009. **8**(2): p. 111-28.

124. Clemente JC, Govindasamy L, Madabushi A, Fisher SZ, Moose RE, Yowell CA, Hidaka K, Kimura T, Hayashi Y, Kiso Y, Agbandje-McKenna M, Dame JB, Dunn BM, and McKenna R. *Structure of the aspartic protease plasmepsin 4 from the malarial parasite Plasmodium malariae bound to an allophenylnorstatine-based inhibitor.* Acta Crystallogr D Biol Crystallogr, 2006. **62**(Pt 3): p. 246-52.

125. Wei D, Jiang X, Zhou L, Chen J, Chen Z, He C, Yang K, Liu Y, Pei J, and Lai L. *Discovery of multitarget inhibitors by combining molecular docking with common pharmacophore matching.* J Med Chem, 2008. **51**(24): p. 7882-8.

126. Ehrman TM, Barlow DJ, and Hylands PJ. *In silico search for multi-target anti-inflammatories in Chinese herbs and formulas.* Bioorg Med Chem, 2010. **18**(6): p. 2204-18.

127. Zhang C, Tan C, Zu X, Zhai X, Liu F, Chu B, Ma X, Chen Y, Gong P, and Jiang Y. *Exploration of (S)-3-aminopyrrolidine as a potentially interesting scaffold for discovery of novel Abl and PI3K dual inhibitors.* European journal of medicinal chemistry, 2011. **46**(4): p. 1404-14.

128. Luan X, Gao C, Zhang N, Chen Y, Sun Q, Tan C, Liu H, Jin Y, and Jiang Y. *Exploration of acridine scaffold as a potentially interesting scaffold for discovering novel multi-target VEGFR-2 and Src kinase inhibitors.* Bioorganic & medicinal chemistry, 2011. **19**(11): p. 3312-9.

129. Li Y, Tan C, Gao C, Zhang C, Luan X, Chen X, Liu H, Chen Y, and Jiang Y. *Discovery of benzimidazole derivatives as novel multi-target EGFR, VEGFR-2 and PDGFR kinase inhibitors.* Bioorganic & medicinal chemistry, 2011. **19**(15): p. 4529-35.

130. Martin EJ and Sullivan DC. *AutoShim: empirically corrected scoring functions for quantitative docking with a crystal structure and IC50 training data.* J Chem Inf Model, 2008. **48**(4): p. 861-72.

131. Martin EJ and Sullivan DC. *Surrogate AutoShim: predocking into a universal ensemble kinase receptor for three dimensional activity prediction, very quickly, without a crystal structure.* J Chem Inf Model, 2008. **48**(4): p. 873-81.

132. Mukherjee P and Martin E. *Development of a minimal kinase ensemble receptor (MKER) for surrogate AutoShim.* J Chem Inf Model, 2011. **51**(10): p. 2697-705.

133. Martin E, Mukherjee P, Sullivan D, and Jansen J. *Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity.* J Chem Inf Model, 2011. **51**(8): p. 1942-56.

134. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, and Tropsha A. *Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds.* J Med Chem, 2004. **47**(9): p. 2356-64.

135. Oloff S, Mailman RB, and Tropsha A. *Application of validated QSAR models of D1 dopaminergic antagonists for database mining.* J Med Chem, 2005. **48**(23): p. 7322-32.

136. Peterson YK, Wang XS, Casey PJ, and Tropsha A. *Discovery of geranylgeranyltransferase-I inhibitors with novel scaffolds by the means of quantitative structure-activity relationship modeling, virtual screening, and experimental validation.* J Med Chem, 2009. **52**(14): p. 4210-20.

137. Tang H, Wang XS, Huang XP, Roth BL, Butler KV, Kozikowski AP, Jung M, and Tropsha A. *Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation.* J Chem Inf Model, 2009. **49**(2): p. 461-76.

138. Xia X, Maliski EG, Gallant P, and Rogers D. *Classification of kinase inhibitors using a Bayesian model.* J Med Chem, 2004. **47**(18): p. 4463-70.

139. Sheridan RP, Feuston BP, Maiorov VN, and Kearsley SK. *Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR.* J Chem Inf Comput Sci, 2004. **44**(6): p. 1912-28.

140. Dragos H, Gilles M, and Alexandre V. *Predicting the predictability: a unified approach to the applicability domain problem of QSAR models.* J Chem Inf Model, 2009. **49**(7): p. 1762-76.

141. Michielan L and Moro S. *Pharmaceutical perspectives of nonlinear QSAR strategies.* J Chem Inf Model, 2010. **50**(6): p. 961-78.

142. Girault JA and Greengard P. *The neurobiology of dopamine signaling.* Arch Neurol, 2004. **61**(5): p. 641-4.

143. Lagger G, O'Carroll D, Rembold M, Khier H, Tischler J, Weitzer G, Schuettengruber B, Hauser C, Brunmeir R, Jenuwein T, and Seiser C. *Essential function of histone deacetylase 1 in proliferation control and CDK inhibitor repression.* EMBO J, 2002. **21**(11): p. 2672-81.

144. Eastman RT, Buckner FS, Yokoyama K, Gelb MH, and Van Voorhis WC. *Thematic review series: lipid posttranslational modifications. Fighting parasitic disease by blocking protein farnesylation.* J Lipid Res, 2006. **47**(2): p. 233-40.

145. Zhang FL and Casey PJ. *Protein prenylation: molecular mechanisms and functional consequences.* Annu Rev Biochem, 1996. **65**: p. 241-69.

146. El Oualid F, Cohen LH, van der Marel GA, and Overhand M. *Inhibitors of protein: geranylgeranyl transferases.* Curr Med Chem, 2006. **13**(20): p. 2385-427.

147. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, and Ye J. *Database resources of the National Center for Biotechnology Information.* Nucleic acids research, 2010. **38**(Database issue): p. D5-16.

148. Fowler G. *cql: Flat file database query language*. in *USENIX Winter 1994 Technical Conference*. 1994. San Francisco, California.

149. Michael J. *Kamfonas/Recursive Hierarchies: The Relational Taboo!* The Relation Journal, 1992.

150. Haughey T. *Modeling Hierarchies*, in *DebTech*2005: Long Branch, New Jersey.

151. Codd EF. *A relational model of data for large shared data banks*, in *Communications of the ACM archive*1970. p. 377-387.

152. *System R did not convince IBM management to abandon its existing product*, in *Funding a Revolution: Government Support for Computing Research*. 1999, National Academies Press.

153. Atwood T. *An Object-Oriented DBMS for Design Support Applications*. in *IEEE COMPINT'85*. 1985. Montréal, Québec, Canada.

154. Derrett N, Kent W, and Lyngbaek P. *Some Aspects of Operations in an Object-Oriented Database.* Database Engineering, 1985. **8**.

155. Maier D, Otis A, and Purdy A. *Object-Oriented Database Development at Servio Logic.* Database Engineering, 1985. **18**.

156. Rao J. *Reasoning about probabilistic parallel programs.* ACM Transactions on Programming Languages and Systems, 1994. **16**(3): p. 798-842.

157. Tropsha A. *Best Practices for QSAR Model Development, Validation, and Exploitation.* Mol. Inf., 2010. **29**: p. 476-488.

158. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, and Overington JP. *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1100-7.

159. Liu T, Lin Y, Wen X, Jorissen RN, and Gilson MK. *BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.* Nucleic Acids Res, 2007. **35**(Database issue): p. D198-201.

160. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, and Bryant SH. *PubChem: a public information system for analyzing bioactivities of small molecules.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W623-33.

161. Irwin JJ and Shoichet BK. *ZINC--a free database of commercially available compounds for virtual screening.* J Chem Inf Model, 2005. **45**(1): p. 177-82.

162. Sadowski J, Gasteiger J, and Klebe G. *Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures.* J. Chem. Inf. Comput. Sci., 1994. **34**: p. 1000-1008.

163. Scior T, Medina-Franco JL, Do QT, Martinez-Mayorga K, Yunes Rojas JA, and Bernard P. *How to recognize and workaround pitfalls in QSAR studies: a critical review.* Curr Med Chem, 2009. **16**(32): p. 4297-313.

164. Han LY, Ma XH, Lin HH, Jia J, Zhu F, Xue Y, Li ZR, Cao ZW, Ji ZL, and Chen YZ. *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, 2008. **26**(8): p. 1276-86.

165. Jorissen RN and Gilson MK. *Virtual screening of molecular databases using a support vector machine.* J. Chem. Inf. Model, 2005. **45**(3): p. 549-61.

166. Glick M, Jenkins JL, Nettles JH, Hitchings H, and Davies JW. *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, 2006. **46**(1): p. 193-200.

167. Lepp Z, Kinoshita T, and Chuman H. *Screening for new antidepressant leads of multiple activities by support vector machines.* J. Chem. Inf. Model, 2006. **46**(1): p. 158-67.

168. Chen B, Harrison RF, Papadatos G, Willett P, Wood DJ, Lewell XQ, Greenidge P, and Stiefl N. *Evaluation of machine-learning methods for ligand-based virtual screening.* J. Comput. Aided Mol. Des., 2007. **21**(1-3): p. 53-62.

169. Franke L, Byvatov E, Werz O, Steinhilber D, Schneider P, and Schneider G. *Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors.* J. Med. Chem, 2005. **48**(22): p. 6997-7004.

170. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, and Schuffenhauer A. *New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching.* J. Chem. Inf. Model, 2006. **46**(2): p. 462-70.

171. Harper G, Bradshaw J, Gittins JC, Green DV, and Leach AR. *Prediction of biological activity for high-throughput screening using binary kernel discrimination.* J. Chem. Inf. Comput. Sci, 2001. **41**(5): p. 1295-300.

172. J. Cui LYH, H.H. Lin, H.L. Zhang, Z.Q. Tang, C.J. Zheng, Z.W. Cao, and Y.Z. Chen. *Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties.* Mol. Immunol, 2007. **44**: p. 866-877.

173. Cai CZ, Han LY, Ji ZL, Chen X, and Chen YZ. *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence.* Nucleic Acids Res., 2003. **31**(13): p. 3692-7.

174. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, and Chen YZ. *Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach.* Nucleic Acids Res., 2004. **32**(21): p. 6437-44.

175. Lin HH, Han LY, Cai CZ, Ji ZL, and Chen YZ. *Prediction of transporter family from protein sequence by support vector machine approach.* Proteins, 2006. **62**(1): p. 218-31.

176. Han LY, Ma XH, Lin HH, Jia J, Zhu F, Xue Y, Li ZR, Cao ZW, Ji ZL, and Chen YZ. *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., 2007: p. (accepted).

177. Wegner JK. *JOELib/JOELib2*, 2005: Department of Computer Science,University of Tübingen: Germany.

178. Han LY, Zheng CJ, Xie B, Jia J, Ma XH, Zhu F, Lin HH, Chen X, and Chen YZ. *Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness.* Drug Discov. Today, 2007. **12**(7-8): p. 304-13.

179. Bocker A, Schneider G, and Teckentrup A. *NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening.* J. Chem. Inf. Model, 2006. **46**(6): p. 2220-9.

180. Oprea TI and Gottfries J. *Chemography: the art of navigating in chemical space.* J. Comb. Chem, 2001. **3**(2): p. 157-66.

181. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, and Chen YZ. *Prediction of P-glycoprotein substrates by a support vector machine approach.* J. Chem. Inf. Comput. Sci, 2004. **44**(4): p. 1497-505.

182. Reymond TFaJ-L. *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., 2007. **(published on Web 01/30/2007)**.

183. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, and Waldmann H. *Charting biologically relevant chemical space: a structural classification of natural products (SCONP).* Proc. Natl. Acad. Sci. U.S.A., 2005. **102**(48): p. 17272-7.

184. Fang H, Tong W, Shi LM, Blair R, Perkins R, Branham W, Hass BS, Xie Q, Dial SL, Moland CL, and Sheehan DM. *Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens.* Chemical Research in Toxicology, 2001. **14**: p. 280-294.

185. Tong W, Xie Q, Hong H, Shi L, Fang H, and Perkins R. *Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity.* Environmental Health Perspectives, 2004. **112**(12): p. 1249-1254.

186. Hu JY and Aizawa T. *Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals.* Water Research, 2003. **37**(6): p. 1213-1222.

187. Jacobs MN. *In silico tools to aid risk assessment of endocrine disrupting chemicals.* Toxicology, 2004. **205**(1-2): p. 43-53.

188. Byvatov E, Fechner U, Sadowski J, and Schneider G. *Comparison of support vector machine and artificial neural network systems for drug/nondrug classification.* Journal of Chemical Information and Computer Sciences, 2003. **43**(6): p. 1882-1889.

189. Doniger S, Hofman T, and Yeh J. *Predicting CNS Permeability of Drug Molecules:Comparison of Neural Network and Support Vector Machine Algorithms.* Journal of Computational Biology, 2002. **9**(6): p. 849-864.

190. He L, Jurs PC, Custer LL, Durham SK, and Pearl GM. *Predicting the Genotoxicity of Polycyclic Aromatic Compounds from Molecular Structure with Different Classifiers.* Chemical Research in Toxicology, 2003. **16**(12): p. 1567-1580.

191. Snyder RD, Pearl GS, Mandakas G, Choy WN, Goodsaid F, and Rosenblum IY. *Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules.* Environmental and Molecular Mutagenesis, 2004. **43**(3): p. 143-158.

192. Grosios K, Wood J, Esser R, Raychaudhuri A, and Dawson J. *Angiogenesis inhibition by the novel VEGF receptor tyrosine kinase inhibitor, PTK787/ZK222584, causes significant anti-arthritic effects in models of rheumatoid arthritis.* Inflamm Res, 2004. **53**(4): p. 133-42.

193. Hall LH KG, Haney DN. *Molconn-Z.* 2002: eduSoft LC: Ashland VA.

194. Li ZR, Han LY, Xue Y, Yap CW, Li H, Jiang L, and Chen YZ. *MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds.* Biotechnology and Bioengineering, 2007. **97**(2): p. 389-396.

195. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, and Willighagen E. *The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics.* J Chem Inf Comput Sci, 2003. **43**(2): p. 493-500.

196. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, and Willighagen EL. *Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics.* Curr Pharm Des, 2006. **12**(17): p. 2111-20.

197. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, and Chen YZ. *Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents.* J Chem Inf Comput Sci, 2004. **44**(5): p. 1630-8.

198. Hemmer MC, Steinhauer V, and Gasteiger J. *Deriving the 3D structure of organic molecules from their infrared spectra.* Vibrational Spectroscopy, 1999. **19**(1): p. 151-164.

199. Rücker G and Rücker C. *Counts of all walks as atomic and molecular descriptors.* Journal of Chemical Information and Computer Sciences, 1993. **33**(5): p. 683-695.

200. Schuur JH, Setzer P, and Gasteiger J. *The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity.* Journal of Chemical Information and Computer Sciences, 1996. **36**(2): p. 334-344.

201. Pearlman RS and Smith KM. *Metric validation and the receptor-relevant subspace concept.* Journal of Chemical Information and Computer Sciences, 1999. **39**(1): p. 28-35.

202. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, and Zaliani A. *MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids.* Journal of Computer-Aided Molecular Design, 1997. **11**(1): p. 79-92.

203. Galvez J, Garcia R, Salabert MT, and Soler R. *Charge indexes. New topological descriptors.* Journal of Chemical Information and Computer Sciences, 1994. **34**(3): p. 520-525.

204. Consonni V, Todeschini R, and Pavan M. *Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors.* Journal of Chemical Information and Computer Sciences, 2002. **42**(3): p. 682-692.

205. Randic M. *Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons.* Tetrahedron, 1975. **31**(11-12): p. 1477-1481.

206. Randic M. *Molecular profiles. Novel geometry-dependent molecular descriptors.* New Journal of Chemistry, 1995. **19**: p. 781-791.

207. Kier LB and Hall LH. *Molecular structure description: The electrotopological state*. 1999, San Diego: Academic Press.

208. Platts JA, Butina D, Abraham MH, and Hersey A. *Estimation of molecular free energy relation descriptors using a group contribution approach.* Journal of Chemical Information and Computer Sciences, 1999. **39**(5): p. 835-845.

209. Livingstone DJ. *Data analysis for chemists: Applications to QSAR and chemical product design*. 1995, Oxford: Oxford University Press.

210. Eriksson L, Johansson E, Kettaneh-Wold N, and Wade KM. *Multi- and megavariate data analysis - Principles and applications*. 2001, Umea, Sweden: Umetrics, AB.

211. Vapnik VN. *An overview of statistical learning theory.* IEEE Trans Neural Netw, 1999. **10**(5): p. 988-99.

212. Vapnik VN. *The nature of statistical learning theory*. 1995, New York: Springer.

213. Burges CJC. *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 1998. **2**(2): p. 127-167.

214. Trotter MWB, Buxton BF, and Holden SB. *Support vector machines in combinatorial chemistry.* Meas. Control, 2001. **34**(8): p. 235-239.

215. Burbidge R, Trotter M, Buxton B, and Holden S. *Drug design by machine learning: support vector machines for pharmaceutical data analysis.* Comput. Chem., 2001. **26**(1): p. 5-14.

216. Czerminski R, Yasri A, and Hartsough D. *Use of support vector machine in pattern classification: Application to QSAR studies.* Quantitative Structure-Activity Relationships, 2001. **20**(3): p. 227-240.

217. Willett P, Barnard JM, and Downs GM. *Chemical Similarity Searching.* J. Chem. Inf. Comput. Sci., 1998. **38**(6): p. 983-996.

218. Bostrom J, Hogner A, and Schmitt S. *Do structurally similar ligands bind in a similar fashion?* J. Med. Chem, 2006. **49**(23): p. 6716-25.

219. Huang N, Shoichet BK, and Irwin JJ. *Benchmarking sets for molecular docking.* J. Med. Chem, 2006. **49**(23): p. 6789-801.

220. Golbraikh A and Tropsha A. *Beware of q2!* J Mol Graph Model, 2002. **20**(4): p. 269-76.

221. Li H, Yap CW, Ung CY, Xue Y, Li ZR, Han LY, Lin HH, and Chen YZ. *Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins.* J Pharm Sci, 2007. **(Published Online)**.

222. Hawkins DM. *The problem of overfitting.* J Chem Inf Comput Sci, 2004. **44**(1): p. 1-12.

223. Downward J. *The ins and outs of signalling.* Nature, 2001. **411**(6839): p. 759-62.

224. McClean MN, Mody A, Broach JR, and Ramanathan S. *Cross-talk and decision making in MAP kinase pathways.* Nat Genet, 2007. **39**(3): p. 409-14.

225. Muller R. *Crosstalk of oncogenic and prostanoid signaling pathways.* J Cancer Res Clin Oncol, 2004. **130**(8): p. 429-44.

226. Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger DA, and Yaffe MB. *A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis.* Science, 2005. **310**(5754): p. 1646-53.

227. Janes KA, Albeck JG, Peng LX, Sorger PK, Lauffenburger DA, and Yaffe MB. *A high-throughput quantitative multiplex kinase assay for monitoring information flow in signaling networks: application to sepsis-apoptosis.* Mol Cell Proteomics, 2003. **2**(7): p. 463-73.

228. Dotto GP. *Crosstalk of Notch with p53 and p63 in cancer growth control.* Nat Rev Cancer, 2009. **9**(8): p. 587-95.

229. Li H and Richardson WD. *Genetics meets epigenetics: HDACs and Wnt signaling in myelin development and regeneration.* Nat Neurosci, 2009. **12**(7): p. 815-7.

230. Zhou W, Feng X, Wu Y, Benge J, Zhang Z, and Chen Z. *FGF-receptor substrate 2 functions as a molecular sensor integrating external regulatory signals into the FGF pathway.* Cell Res, 2009.

231. Adams JM and Cory S. *The Bcl-2 apoptotic switch in cancer development and therapy.* Oncogene, 2007. **26**(9): p. 1324-37.

232. Buchanan FG and DuBois RN. *Connecting COX-2 and Wnt in cancer.* Cancer Cell, 2006. **9**(1): p. 6-8.

233. Castellone MD, Teramoto H, Williams BO, Druey KM, and Gutkind JS. *Prostaglandin E2 promotes colon cancer cell growth through a Gs-axin-beta-catenin signaling axis.* Science, 2005. **310**(5753): p. 1504-10.

234. Eibl G. *The Role of PPAR-gamma and Its Interaction with COX-2 in Pancreatic Cancer.* PPAR Res, 2008. **2008**: p. 326915.

235. Numakawa T, Kumamaru E, Adachi N, Yagasaki Y, Izumi A, and Kunugi H. *Glucocorticoid receptor interaction with TrkB promotes BDNF-triggered PLC-gamma signaling for glutamate release via a glutamate transporter.* Proc Natl Acad Sci U S A, 2009. **106**(2): p. 647-52.

236. Schwappacher R, Weiske J, Heining E, Ezerski V, Marom B, Henis YI, Huber O, and Knaus P. *Novel crosstalk to BMP signalling: cGMP-dependent kinase I modulates BMP receptor and Smad activity.* Embo J, 2009. **28**(11): p. 1537-50.

237. Thomas SM, Bhola NE, Zhang Q, Contrucci SC, Wentzel AL, Freilino ML, Gooding WE, Siegfried JM, Chan DC, and Grandis JR. *Cross-talk between G protein-coupled receptor and epidermal growth factor receptor signaling pathways contributes to growth and invasion of head and neck squamous cell carcinoma.* Cancer Res, 2006. **66**(24): p. 11831-9.

238. Scagliotti GV. *Potential role of multi-targeted tyrosine kinase inhibitors in non-small-cell lung cancer.* Ann Oncol, 2007. **18 Suppl 10**: p. x32-41.

239. Zimmermann GR, Lehar J, and Keith CT. *Multi-target therapeutics: when the whole is greater than the sum of the parts.* Drug Discov Today, 2007. **12**(1-2): p. 34-42.

240. C VS, Babar SM, Song EJ, Oh E, and Yoo YS. *Kinetic analysis of the MAPK and PI3K/Akt signaling pathways.* Mol Cells, 2008. **25**(3): p. 397-406.

241. Papin JA and Palsson BO. *The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis.* Biophys J, 2004. **87**(1): p. 37-46.

242. Kanehisa M, Goto S, Kawashima S, Okuno Y, and Hattori M. *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32**(Database issue): p. D277-80.

243. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, and Pellegrini-Toole A. *The EcoCyc and MetaCyc databases.* Nucleic Acids Res, 2000. **28**(1): p. 56-9.

244. Urbanczyk-Wochniak E and Sumner LW. *MedicCyc: a biochemical pathway database for Medicago truncatula.* Bioinformatics, 2007. **23**(11): p. 1418-23.

245. Mueller LA, Zhang P, and Rhee SY. *AraCyc: a biochemical pathway database for Arabidopsis.* Plant Physiol, 2003. **132**(2): p. 453-60.

246. Koike A, Kobayashi Y, and Takagi T. *Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource.* Genome Res, 2003. **13**(6A): p. 1231-43.

247. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, and Bork P. *STRING 7--recent developments in the integration and prediction of protein interactions.* Nucleic Acids Res, 2007. **35**(Database issue): p. D358-62.

248. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, and Stein L. *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.

249. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, and Yaschenko E. *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2008. **36**(Database issue): p. D13-21.

250. UniProt C. *The universal protein resource (UniProt).* Nucleic Acids Res, 2008. **36**(Database issue): p. D190-5.

251. Ambion. *Ambion, Inc. - The RNA Company*. Available from: http://www.ambion.com/.

252. Signaling C. *Cell Signaling Technology*. Available from: http://www.cellsignal.com/.

253. Chen X, Ji ZL, and Chen YZ. *TTD: Therapeutic Target Database.* Nucleic Acids Res, 2002. **30**(1): p. 412-5.

254. Ji ZL, Han LY, Yap CW, Sun LZ, Chen X, and Chen YZ. *Drug Adverse Reaction Target Database (DART) : proteins related to adverse drug reactions.* Drug Saf, 2003. **26**(10): p. 685-90.

255. Ji ZL, Sun LZ, Chen X, Zheng CJ, Yao LX, Han LY, Cao ZW, Wang JF, Yeo WK, Cai CZ, and Chen YZ. *Internet resources for proteins associated with drug therapeutic effects, adverse reactions and ADME.* Drug Discov Today, 2003. **8**(12): p. 526-9.

256. Sun LZ, Ji ZL, Chen X, Wang JF, and Chen YZ. *ADME-AP: a database of ADME associated proteins.* Bioinformatics, 2002. **18**(12): p. 1699-700.

257. Zheng CJ, Han LY, Xie B, Liew CY, Ong S, Cui J, Zhang HL, Tang ZQ, Gan SH, Jiang L, and Chen YZ. *PharmGED: Pharmacogenetic Effect Database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D794-9.

258. BioCarta. *BioCarta - Charting Pathways of Life*. Available from: http://www.biocarta.com/.

259. Lyons SA, Chung WJ, Weaver AK, Ogunrinu T, and Sontheimer H. *Autocrine glutamate signaling promotes glioma cell invasion.* Cancer Res, 2007. **67**(19): p. 9463-71.

260. Morales M, Gonzalez-Mejia ME, Bernabe A, Hernandez-Kelly LC, and Ortega A. *Glutamate activates protein kinase B (PKB/Akt) through AMPA receptors in cultured Bergmann glia cells.* Neurochem Res, 2006. **31**(3): p. 423-9.

261. Ishiuchi S, Yoshida Y, Sugawara K, Aihara M, Ohtani T, Watanabe T, Saito N, Tsuzuki K, Okado H, Miwa A, Nakazato Y, and Ozawa S. *Ca2+-permeable AMPA receptors regulate growth of human glioblastoma via Akt activation.* J Neurosci, 2007. **27**(30): p. 7987-8001.

262. Argiris A, Wang CX, Whalen SG, and DiGiovanna MP. *Synergistic interactions between tamoxifen and trastuzumab (Herceptin).* Clin Cancer Res, 2004. **10**(4): p. 1409-20.

263. Osborne CK and Schiff R. *Growth factor receptor cross-talk with estrogen receptor as a mechanism for tamoxifen resistance in breast cancer.* Breast, 2003. **12**(6): p. 362-7.

264. Coradini D, Biffi A, Cappelletti V, and Di Fronzo G. *Activity of tamoxifen and new antiestrogens on estrogen receptor positive and negative breast cancer cells.* Anticancer Res, 1994. **14**(3A): p. 1059-64.

265. Nahta R and Esteva FJ. *Trastuzumab: triumphs and tribulations.* Oncogene, 2007. **26**(25): p. 3637-43.

266. Osborne CK, Shou J, Massarweh S, and Schiff R. *Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer.* Clin Cancer Res, 2005. **11**(2 Pt 2): p. 865s-70s.

267. Schiff R, Massarweh SA, Shou J, Bharwani L, Arpino G, Rimawi M, and Osborne CK. *Advanced concepts in estrogen receptor biology and breast cancer endocrine resistance: implicated role of growth factor signaling and estrogen receptor coregulators.* Cancer Chemother Pharmacol, 2005. **56 Suppl 1**: p. 10-20.

268. Glaros S, Atanaskova N, Zhao C, Skafar DF, and Reddy KB. *Activation function-1 domain of estrogen receptor regulates the agonistic and antagonistic actions of tamoxifen.* Mol Endocrinol, 2006. **20**(5): p. 996-1008.

269. Massarweh S and Schiff R. *Resistance to endocrine therapy in breast cancer: exploiting estrogen receptor/growth factor signaling crosstalk.* Endocr Relat Cancer, 2006. **13 Suppl 1**: p. S15-24.

270. Jemal A, Siegel R, Ward E, Murray T, Xu J, and Thun MJ. *Cancer statistics, 2007.* CA Cancer J Clin, 2007. **57**(1): p. 43-66.

271. Society AC. *Cancer Facts & Figures 2012*, 2012.

272. Thomson S, Petti F, Sujka-Kwok I, Epstein D, and Haley JD. *Kinase switching in mesenchymal-like non-small cell lung cancer lines contributes to EGFR inhibitor resistance through pathway redundancy.* Clin Exp Metastasis, 2008. **25**(8): p. 843-54.

273. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, and Ye J. *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2012. **40**(Database issue): p. D13-25.

274. Walker T, Grulke CM, Pozefsky D, and Tropsha A. *Chembench: a cheminformatics workbench.* Bioinformatics, 2010. **26**(23): p. 3000-1.

275. Fechner N, Jahn A, Hinselmann G, and Zell A. *Estimation of the applicability domain of kernel-based machine learning models for virtual screening.* J Cheminform, 2010. **2**(1): p. 2.

276. Ma XH, Wang R, Yang SY, Li ZR, Xue Y, Wei YC, Low BC, and Chen YZ. *Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds.* J Chem Inf Model, 2008. **48**(6): p. 1227-37.

277. Fink T and Reymond JL. *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, 2007. **47**(2): p. 342-53.

278. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, and Waldmann H. *Charting biologically relevant chemical space: a structural classification of natural products (SCONP).* Proc Natl Acad Sci U S A, 2005. **102**(48): p. 17272-7.

279. Yap CW. *PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints.* J Comput Chem, 2011. **32**(7): p. 1466-74.

280. Chekmarev D, Kholodovych V, Kortagere S, Welsh WJ, and Ekins S. *Predicting inhibitors of acetylcholinesterase by regression and classification machine learning approaches with combinations of molecular descriptors.* Pharm Res, 2009. **26**(9): p. 2216-24.

281. Sun M, Chen J, Wei H, Yin S, Yang Y, and Ji M. *Quantitative structure-activity relationship and classification analysis of diaryl ureas against vascular endothelial growth factor receptor-2 kinase using linear and non-linear models.* Chem Biol Drug Des, 2009. **73**(6): p. 644-54.

282. Fruitet J, McFarland DJ, and Wolpaw JR. *A comparison of regression techniques for a two-dimensional sensorimotor rhythm-based brain-computer interface.* J Neural Eng, 2010. **7**(1): p. 16003.

283. Balabin RM and Lomakina EI. *Support vector machine regression (SVR/LS-SVM)--an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data.* Analyst, 2011. **136**(8): p. 1703-12.

284. Li L, Wang B, and Meroueh SO. *Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries.* J Chem Inf Model, 2011. **51**(9): p. 2132-8.

285. Devillers J, Doucet JP, Doucet-Panaye A, Decourtye A, and Aupinel P. *Linear and non-linear QSAR modelling of juvenile hormone esterase inhibitors.* SAR QSAR Environ Res, 2012. **23**(3-4): p. 357-69.

286. Li F and Yang Y. *Analysis of recursive gene selection approaches from microarray data.* Bioinformatics, 2005. **21**(19): p. 3741-7.

287. Pochet N, De Smet F, Suykens JA, and De Moor BL. *Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction.* Bioinformatics, 2004. **20**(17): p. 3185-95.

288. Chang CC and Lin CJ. *LIBSVM: A library for support vector machines.* ACM TIST, 2011. **2**(3): p. 27:1--27:27.

289. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, and Head MS. *A critical assessment of docking programs and scoring functions.* J Med Chem, 2006. **49**(20): p. 5912-31.

290. Pilpel Y, Sudarsanam P, and Church GM. *Identifying regulatory networks by combinatorial analysis of promoter elements.* Nat Genet, 2001. **29**(2): p. 153-9.

291. Sergina NV, Rausch M, Wang D, Blair J, Hann B, Shokat KM, and Moasser MM. *Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3.* Nature, 2007. **445**(7126): p. 437-41.

292. Christopher M., Overall, and Kleifeld O. *Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy.* Nature Reviews Cancer 2006. **6**: p. 227-239.

293. Force T, Krause DS, and Van Etten RA. *Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition.* Nat Rev Cancer, 2007. **7**(5): p. 332-44.

294. Krug M and Hilgeroth A. *Recent advances in the development of multi-kinase inhibitors.* Mini Rev Med Chem, 2008. **8**(13): p. 1312-27.

295. Shoichet BK. *Virtual screening of chemical libraries.* Nature, 2004. **432**(7019): p. 862-5.

296. Yamane S, Ishida S, Hanamoto Y, Kumagai K, Masuda R, Tanaka K, Shiobara N, Yamane N, Mori T, Juji T, Fukui N, Itoh T, Ochi T, and Suzuki R. *Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients.* J Inflamm (Lond), 2008. **5**: p. 5.

297. Deng XQ, Wang HY, Zhao YL, Xiang ML, Jiang PD, Cao ZX, Zheng YZ, Luo SD, Yu LT, Wei YQ, and Yang SY. *Pharmacophore modelling and virtual screening for identification of new Aurora-A kinase inhibitors.* Chem Biol Drug Des, 2008. **71**(6): p. 533-9.

298. Deanda F, Stewart EL, Reno MJ, and Drewry DH. *Kinase-Targeted Library Design through the Application of the PharmPrint Methodology.* J Chem Inf Model, 2008. **48**(12): p. 2395-403.

299. Briem H and Gunther J. *Classifying "kinase inhibitor-likeness" by using machine-learning methods.* Chembiochem, 2005. **6**(3): p. 558-66.

300. Gundla R, Kazemi R, Sanam R, Muttineni R, Sarma JA, Dayam R, and Neamati N. *Discovery of novel small-molecule inhibitors of human epidermal growth factor receptor-2: combined ligand and target-based approach.* J Med Chem, 2008. **51**(12): p. 3367-77.

301. Prado-Prado FJ, de la Vega OM, Uriarte E, Ubeira FM, Chou KC, and Gonzalez-Diaz H. *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, 2008.

302. Zhang X and Fernandez A. *In silico drug profiling of the human kinome based on a molecular marker for cross reactivity.* Mol Pharm, 2008. **5**(5): p. 728-38.

303. Gockel I, Moehler M, Frerichs K, Drescher D, Trinh TT, Duenschede F, Borschitz T, Schimanski K, Biesterfeld S, Herzer K, Galle PR, Lang H, Junginger T, and Schimanski CC. *Co-expression of receptor tyrosine kinases in esophageal adenocarcinoma and squamous cell cancer.* Oncol Rep, 2008. **20**(4): p. 845-50.

304. Stommel JM, Kimmelman AC, Ying H, Nabioullin R, Ponugoti AH, Wiedemeyer R, Stegh AH, Bradner JE, Ligon KL, Brennan C, Chin L, and DePinho RA. *Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies.* Science, 2007. **318**(5848): p. 287-90.

305. Speake G, Holloway B, and Costello G. *Recent developments related to the EGFR as a target for cancer chemotherapy.* Curr Opin Pharmacol, 2005. **5**(4): p. 343-9.

306. Moasser MM. *Targeting the function of the HER2 oncogene in human cancer therapeutics.* Oncogene, 2007. **26**(46): p. 6577-92.

307. Zhong H and Bowen JP. *Molecular design and clinical development of VEGFR kinase inhibitors.* Curr Top Med Chem, 2007. **7**(14): p. 1379-93.

308. Lewis NL. *The platelet-derived growth factor receptor as a therapeutic target.* Curr Oncol Rep, 2007. **9**(2): p. 89-95.

309. Rusnati M and Presta M. *Fibroblast growth factors/fibroblast growth factor receptors as targets for the development of anti-angiogenesis strategies.* Curr Pharm Des, 2007. **13**(20): p. 2025-44.

310. Benati D and Baldari CT. *SRC family kinases as potential therapeutic targets for malignancies and immunological disorders.* Curr Med Chem, 2008. **15**(12): p. 1154-65.

311. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, and Watson P. *Virtual screening using protein-ligand docking: avoiding artificial enrichment.* J Chem Inf Comput Sci, 2004. **44**(3): p. 793-806.

312. Glick M, Jenkins JL, Nettles JH, Hitchings H, and Davies JW. *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, 2006. **46**(1): p. 193-200.

313. Ma XH, Jia J, Zhu F, Xue Y, Li ZR, and Chen YZ. *Comparative Analysis of Machine Learning Methods in Ligand Based Virtual Screening of Large Compound Libraries.* Comb. Chem. High Throughput Screen, 2009: p. (accepted).

314. Carvalho JF, Blank M, and Shoenfeld Y. *Vascular endothelial growth factor (VEGF) in autoimmune diseases.* J Clin Immunol, 2007. **27**(3): p. 246-56.

315. Daouti S, Latario B, Nagulapalli S, Buxton F, Uziel-Fusi S, Chirn GW, Bodian D, Song C, Labow M, Lotz M, Quintavalla J, and Kumar C. *Development of comprehensive functional genomic screens to identify novel mediators of osteoarthritis.* Osteoarthritis Cartilage, 2005. **13**(6): p. 508-18.

316. Remmers EF, Sano H, and Wilder RL. *Platelet-derived growth factors and heparin-binding (fibroblast) growth factors in the synovial tissue pathology of rheumatoid arthritis.* Semin Arthritis Rheum, 1991. **21**(3): p. 191-9.