molecular pharmaceutics

Virtual Screening of Selective Multitarget Kinase Inhibitors by Combinatorial Support Vector Machines

X. H. Ma,[†,‡] R. Wang,[†] C. Y. Tan,[§] Y. Y. Jiang,[§] T. Lu,[||] H. B. Rao,[||] X. Y. Li,[||] M. L. Go,[⊥] B. C. Low,[‡] and Y. Z. Chen*[,†,§,||]

Bioinformatics and Drug Design Group, Department of Pharmacy, Centre for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, The Key Laboratory of Chemical Biology, Guangdong Province, The Graduate School at Shenzhen, Tsinghua University, Shenzhen, Guangdong, P. R. China, College of Chemical Engineering and State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610065, P. R. China, Department of Pharmacy, National University of Singapore, Science Drive 4, Singapore 117543, and Department of Biological Science, National University of Singapore, Blk S2, Level 5, Science Drive 4, Singapore 117543

Received August 31, 2009; Revised Manuscript Received August 7, 2010; Accepted August 16, 2010

Abstract: Multitarget agents have been increasingly explored for enhancing efficacy and reducing countertarget activities and toxicities. Efficient virtual screening (VS) tools for searching selective multitarget agents are desired. Combinatorial support vector machines (C-SVM) were tested as VS tools for searching dual-inhibitors of 11 combinations of 9 anticancer kinase targets (EGFR, VEGFR, PDGFR, Src, FGFR, Lck, CDK1, CDK2, GSK3). C-SVM trained on 233−1,316 non-dual-inhibitors correctly identified 26.8%−57.3% (majority >36%) of the 56−230 intra-kinase-group dual-inhibitors (equivalent to the 50−70% yields of two independent individual target VS tools), and 12.2% of the 41 inter-kinase-group dual-inhibitors. C-SVM were fairly selective in misidentifying as dual-inhibitors 3.7%−48.1% (majority <20%) of the 233−1,316 non-dual-inhibitors of the same kinase pairs and 0.98%−4.77% of the 3,971−5,180 inhibitors of other kinases. C-SVM produced low false-hit rates in misidentifying as dual-inhibitors 1,746−4,817 (0.013%−0.036%) of the 13.56 M PubChem compounds, 12−175 (0.007%−0.104%) of the 168 K MDDR compounds, and 0−84 (0.0%−2.9%) of the 19,495−38,483 MDDR compounds similar to the known dual-inhibitors. C-SVM was compared to other VS methods Surflex-Dock, DOCK Blaster, kNN and PNN against the same sets of kinase inhibitors and the full set or subset of the 1.02 M Zinc clean-leads data set. C-SVM produced comparable dual-inhibitor yields, slightly better false-hit rates for kinase inhibitors, and significantly lower false-hit rates for the Zinc clean-leads data set. Combinatorial SVM showed promising potential for searching selective multitarget agents against intra-kinase-group kinases without explicit knowledge of multitarget agents.

Keywords: Anticancer; multitarget; high-throughput screening; computer aided drug design; kinase inhibitor; support vector machines; virtual screening

Introduction

A 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,[1] redundancy,[2] crosstalk,[3] com-

* Corresponding author. Mailing address: National University of Singapore, Department of Pharmacy, Blk S16, Level 8, Science Drive 2, Singapore 117543. Tel: 65-6874-6877. Fax: 65-6774-6756. E-mail: phacyz@nus.edu.sg.

[†] Bioinformatics and Drug Design Group, Department of Pharmacy, Centre for Computational Science and Engineering, National University of Singapore.

[‡] Department of Biological Science, National University of Singapore.

[§] Tsinghua University.

[||] Sichuan University.

[⊥] Department of Pharmacy, National University of Singapore.

[1] Smalley, K. S.; Haass, N. K.; Brafford, P. A.; Lioni, M.; Flaherty, K. T.; 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), 1136–44.

[2] Pilpel, Y.; Sudarsanam, P.; Church, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **2001**, *29* (2), 153–9.

10.1021/mp100179t © 2010 American Chemical Society
Published on Web 08/16/2010

pensatory and neutralizing actions,[4] antitarget and countertarget activities,[5] and on-target and off-target toxicities.[6] Multitarget agents and drug combinations have been increasingly explored[1,7] for enhancing therapeutic efficacies and improving safety and resistance profiles by selectively modulating the elements of these countertarget and toxicity activities.[8] In particular, multitarget 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.[9,10] Methods for efficient search of multitarget agents are highly desired.

Virtual screening (VS) methods have been widely explored for facilitating lead discovery against individual targets.[11−13] In particular, molecular docking,[14] pharmacophore,[15] QSAR,[16] machine learning,[17] and combination methods[18] have been extensively used for VS of single-target kinase inhibitors, but few multitarget VS studies have been reported.[19,20] An interesting strategy for identifying multitarget kinase inhibitors is to use experimentally obtained small-scale profiles for predicting inhibitors of a larger kinase set.[20] In principle, single-target VS tools may be combined to collectively identify multitarget 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 multitarget agents that are significantly fewer in numbers and more sparsely distributed in the chemical space than non-dual-inhibitors (Table 1).

An extensively used machine learning method, support vector machines (SVM), may be potentially explored as multitarget VS tools because it has shown high yields and low false-hit rates in searching single-target agents[21] sometimes based on sparsely distributed active compounds.[13] SVM, trained based on the physicochemical properties of active and inactive compounds, 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. Multitarget VS performance of combinatorial SVMs (C-SVM), which com-

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

(4) Sergina, N. V.; Rausch, M.; Wang, D.; Blair, J.; Hann, B.; Shokat, K. M.; Moasser, M. M. Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature* **2007**, *445* (7126), 437–41.

(5) Overall, C. M.; Kleifeld, O. Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. *Nat. Rev. Cancer* **2006**, *6*, 227–39.

(6) Force, T.; Krause, D. S.; Van Etten, R. A. Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition. *Nat. Rev. Cancer* **2007**, *7* (5), 332–44.

(7) Keith, C. T.; Borisy, A. A.; Stockwell, B. R. Multicomponent therapeutics for networked systems. *Nat. Rev. Drug Discovery* **2005**, *4* (1), 71–8.

(8) Larder, B. A.; Kemp, S. D.; Harrigan, P. R. Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy. *Science* **1995**, *269* (5224), 696–9.

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

(10) Gill, A. L.; Verdonk, M.; Boyle, R. G.; 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), 1408–22.

(11) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862–5.

(12) 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.; Suzuki, R. Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients. *J. Inflammation (London, U.K.)* **2008**, *5*, 5.

(13) Ma, X. H.; Wang, R.; Yang, S. Y.; Li, Z. R.; Xue, Y.; Wei, Y. C.; Low, B. C.; Chen, Y. Z. Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. *J. Chem. Inf. Model.* **2008**, *48* (6), 1227–37.

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

(15) Deng, X. Q.; Wang, H. Y.; Zhao, Y. L.; Xiang, M. L.; Jiang, P. D.; Cao, Z. X.; Zheng, Y. Z.; Luo, S. D.; Yu, L. T.; Wei, Y. Q.; Yang, S. Y. Pharmacophore modelling and virtual screening for identification of new Aurora-A kinase inhibitors. *Chem. Biol. Drug Des.* **2008**, *71* (6), 533–9.

(16) Deanda, F.; Stewart, E. L.; Reno, M. J.; Drewry, D. H. Kinase-Targeted Library Design through the Application of the PharmPrint Methodology. *J. Chem. Inf. Model.* **2008**, *48* (12), 2395–403.

(17) Briem, H.; Gunther, J. Classifying "kinase inhibitor-likeness" by using machine-learning methods. *ChemBioChem* **2005**, *6* (3), 558–66.

(18) Gundla, R.; Kazemi, R.; Sanam, R.; Muttineni, R.; Sarma, J. A.; Dayam, R.; 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), 3367–77.

(19) Prado-Prado, F. J.; de la Vega, O. M.; Uriarte, E.; Ubeira, F. M.; Chou, K. C.; 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.* **2009**, *17* (2), 569–75.

(20) Zhang, X.; Fernandez, A. In silico drug profiling of the human kinome based on a molecular marker for cross reactivity. *Mol. Pharmaceutics* **2008**, *5* (5), 728–38.

(21) Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z. 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. Graphics Model.* **2008**, *26* (8), 1276–86.

Table 1. Data Sets of Dual-Inhibitors and Non-Dual-Inhibitors of the Kinase Pairs Used for Developing and Testing Combinatorial SVM Dual-Inhibitor Virtual Screening Tools[a]

	inhibitors in training sets					
	training set for kinase A			training set for kinase B		
kinase pair A-B	no. of inhibitors of A that are non-inhibitor of B (no. of families)	no. of these inhibitors that are in the B inhibitor families (no. of families)	no. of these inhibitors that are in the families of dual-inhibitors of A and B (no. of families)	no. of inhibitors of B that are non-inhibitor of A (no. of families)	no. of these inhibitors that are in the A inhibitor families (no. of families)	no. of these inhibitors that are in the families of dual-inhibitors of A and B (no. of families)
EGFR-PDGFR	1316 (384)	336 (70)	100 (19)	622 (202)	251 (70)	153 (23)
EGFR-FGFR	1303 (388)	284 (52)	160 (22)	392 (131)	154 (52)	124 (27)
EGFR-Src	1262 (372)	331 (73)	166 (31)	748 (216)	243 (73)	168 (38)
VEGFR-Lck	1232 (427)	220 (69)	102 (17)	445 (171)	206 (69)	52 (11)
PDGFR-FGFR	450 (168)	100 (29)	118 (27)	233 (90)	89 (29)	79 (25)
PDGFR-Src	492 (174)	237 (53)	144 (24)	672 (213)	206 (53)	170 (38)
Src-Lck	804 (236)	222 (49)	98 (11)	450 (175)	160 (49)	23 (9)
CDK1-CDK2	484 (199)	183 (52)	99 (28)	650 (251)	178 (52)	68 (34)
CDK1-GSK3	503 (224)	140 (45)	38 (20)	642 (266)	143 (45)	83 (22)
CDK2-GSK3	749 (280)	226 (62)	58 (23)	722 (275)	249 (62)	107 (24)
CDK1-VEGFR	651 (251)	250 (75)	23 (8)	1285(434)	251 (75)	70 (17)

	inhibitors and other compounds in testing set					
	dual-inhibitors of A and B					
kinase pair A-B	no. of dual-inhibitors of A and B (no. of families)	no. (%) of dual-inhibitors in the families that contain both A and B non-dual-inhibitor in training sets	no. (%) of dual-inhibitors of A and B as inhibitor of at least one of the other 7 kinases studied in this work	no. (%) of dual-inhibitors of A and B as inhibitor of more than 2 of the other 7 kinases studied in this work	no. of inhibitors of other 7 kinases	no. of MDDR compds similar to dual-inhibitors of A and B
EGFR-PDGFR	58 (40)	22 (37.9)	50 (86.2)	3 (5.2)	4097	3806
EGFR-FGFR	71 (39)	37 (52.1)	70 (98.6)	2 (2.8)	4327	1001
EGFR-Src	112 (64)	46 (41.1)	46 (41.1)	2 (1.8)	3971	1127
VEGFR-Lck	61 (23)	29 (47.5)	37 (60.7)	0 (0.0)	4355	413
PDGFR-FGFR	230 (78)	90 (39.1)	214 (93.0)	3 (1.3)	5180	3614
PDGFR-Src	188 (67)	71 (37.8)	184 (97.9)	3 (1.6)	4741	2893
Src-Lck	56 (17)	23 (41.1)	38 (67.9)	0 (0.0)	4783	276
CDK1-CDK2	174 (84)	53 (30.5)	24 (13.8)	0 (0.0)	4785	2629
CDK1-GSK3	155 (51)	49 (31.6)	17 (11.0)	0 (0.0)	4793	3279
CDK2-GSK3	75 (44)	31 (41.3)	17 (22.7)	0 (0.0)	4547	1617
CDK1-VEGFR	41 (25)	7 (17.1)	0 (0.0)	0 (0.0)	4149	427

[a] Additional sets of 13.56 million PubChem compounds and 168 thousand MDDR active compounds were also used for the test.

bine the prediction of two separate SVM classifiers for each of the multiple kinases, was tested by using them to search dual-inhibitors of combinations of 9 anticancer kinase targets EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3. These kinase targets were selected because of their therapeutic relevance and the availability of sufficient number of the known inhibitors and dual-inhibitors. The first six kinases belong to the protein kinase group PTK group, and the last three belong to the CMGC group respectively.

Based on dual-inhibitor availability, we focused on 11 kinase pairs EGFR-PDGFR, EGFR-FGFR, EGFR-Src, VEGFR-Lck, PDGFR-FGFR, PDGFR-Src, Src-Lck, CDK1-CDK2, CDK1-GSK3, CDK2-GSK3, and CDK1-VEGFR. The first 7 kinase pairs are intra-PTK group, the eighth to 10th are intra-CMGC group, and the 11th are inter-PTK-CMGC group kinase pairs respectively, representative of different types of kinase pairs. These kinase pairs are frequently coexpressed or coactivated in various cancers,[22,23] and targeted by multitarget agents[9,10] with good anticancer efficacies. Inhibitors of growth factor receptor tyrosine kinases EGFR, VEGFR, PDGFR and FGFR have been successfully used for cancer treatments.[10,24−28] EGFR promotes proliferation and survival.[24] VEGFR regulates angiogenesis and survival.[26] PDGFR modulates angiogenesis and growth, and is one of the multitargets of several approved and clinical trial drugs.[10,27] FGFR regulates angiogenesis and cancer progression, and is one of the multitargets of several clinical trial drugs.[10,28] Src-family kinases Src and Lck modulate multiple pathways of cell growth, differentiation, migration and survival, and are part of the multitargets of several marketed and clinical trial drugs.[10,29] CDKs promote cell cycle progression, their inhibition severely limits the aberrant cell-cycle process in tumor and induces apoptosis, and CDK inhibitors are being developed and tested in clinical trials for anticancer therapeutics.[30] GSK3 modulates glucose metabolism and the function of various proteins, and is associated with neurodegenerative diseases, stroke, bipolar disorder, diabetes and cancer.[31] GSK3 inhibitors have started to reach clinical development for the treatment of various disorders.[31]

Multitarget VS performance was tested by a rigorous method that assumes no explicit knowledge of known multitarget agents, because the number of known multitarget agents is generally small for many target pairs. The SVM of each kinase was trained by using non-dual-inhibitors of that kinase. The collective yield of C-SVM 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 C-SVM was assessed by using non-dual-inhibitors of the kinase pair and inhibitors of the other 7 kinases, out of the 9 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 compounds, 168 thousand compounds from the MDL Drug Data Report (MDDR) database, and 276−3,806 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.[32] Therefore, VS performance is more strictly tested by using a subset of MDDR compounds similar to the dual-inhibitors so that enrichment is not simply a separation of trivial physicochemical features.[33]

VS performance of C-SVM was further compared with that of three VS methods, which include two popular molecular docking software Surflex-Dock and DOCK version 3.5.54 at the DOCK Blaster server,[34] a similarity-based

[22] Gockel, I.; Moehler, M.; Frerichs, K.; Drescher, D.; Trinh, T. T.; Duenschede, F.; Borschitz, T.; Schimanski, K.; Biesterfeld, S.; Herzer, K.; Galle, P. R.; Lang, H.; Junginger, T.; Schimanski, C. C. Co-expression of receptor tyrosine kinases in esophageal adenocarcinoma and squamous cell cancer. *Oncol. Rep.* **2008**, *20* (4), 845–50.

[23] Stommel, J. M.; Kimmelman, A. C.; Ying, H.; Nabioullin, R.; Ponugoti, A. H.; Wiedemeyer, R.; Stegh, A. H.; Bradner, J. E.; Ligon, K. L.; Brennan, C.; Chin, L.; DePinho, R. A. Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. *Science* **2007**, *318* (5848), 287–90.

[24] Speake, G.; Holloway, B.; Costello, G. Recent developments related to the EGFR as a target for cancer chemotherapy. *Curr. Opin. Pharmacol.* **2005**, *5* (4), 343–9.

[25] Moasser, M. M. Targeting the function of the HER2 oncogene in human cancer therapeutics. *Oncogene* **2007**, *26* (46), 6577–92.

[26] Zhong, H.; Bowen, J. P. Molecular design and clinical development of VEGFR kinase inhibitors. *Curr. Top. Med. Chem.* **2007**, *7* (14), 1379–93.

[27] Lewis, N. L. The platelet-derived growth factor receptor as a therapeutic target. *Curr. Oncol. Rep.* **2007**, *9* (2), 89–95.

[28] Rusnati, M.; Presta, M. Fibroblast growth factors/fibroblast growth factor receptors as targets for the development of anti-angiogenesis strategies. *Curr. Pharm. Des.* **2007**, *13* (20), 2025–44.

[29] Benati, D.; Baldari, C. T. SRC family kinases as potential therapeutic targets for malignancies and immunological disorders. *Curr. Med. Chem.* **2008**, *15* (12), 1154–65.

[30] Schwartz, G. K.; Shah, M. A. Targeting the cell cycle: a new approach to cancer therapy. *J. Clin. Oncol.* **2005**, *23* (36), 9408–21.

[31] Medina, M.; Castro, A. Glycogen synthase kinase-3 (GSK-3) inhibitors reach the clinic. *Curr. Opin. Drug Discovery Dev.* **2008**, *11* (4), 533–43.

[32] Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793–806.

[33] Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–801.

[34] Irwin, J. J.; Shoichet, B. K.; Mysinger, M. M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Automated docking screens: a feasibility study. *J. Med. Chem.* **2009**, *52* (18), 5712–20.

statistical learning method k nearest neighbor (kNN),[35] and a machine-learning method probabilistic neural networks (PNN)[36] against the same sets of dual- and non-dual kinase inhibitors and the full set or subset of the 1.02 million Zinc clean-leads data set (Zinc-CLD).[37] The specific indicators to be compared are the dual-inhibitor yields for both intragroup and intergroup dual-kinase inhibitors, and the false-hit rates for non-dual kinase inhibitors and the Zinc-CLD data set, which enable objective assessment of the capability of C-SVM with respect to those of the popular as well as machine learning based VS methods.

Methods

Compound Collection, Training and Testing Data Sets, Molecular Descriptors. A total of 233−1,316 non-dual-inhibitors of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3, and 41−230 dual-inhibitors of EGFR-PDGFR, EGFR-FGFR, EGFR-Src, VEGFR-Lck, PDGFR-FGFR, PDGFR-Src, Src-Lck, CDK1-CDK2, CDK1-GSK3, CDK2-GSK3, and CDK1-VEGFR, each with IC50 $\leq$ 10 μM, were collected from the literature[38−47] and the BindingDB database.[48] 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 1 summarizes the statistics of these inhibitors and MDDR compounds similar to at least one dual-inhibitor. Figure 1 shows the Venn graph of our collected dual-inhibitors, the 11 evaluated kinase pairs and non-dual-inhibitors of the 9 evaluated kinases. As few noninhibitors have been reported, putative noninhibitors of each kinase were generated by using our published method that requires no knowledge of inactive compounds or active compounds of other target classes and enables more expanded coverage of the "noninhibitor" chemical space.[12,13] First, 13.56 million PubChem and 168 thousand MDDR compounds were clustered into 8,993 compound families of similar molecular descriptors,[49] which are consistent with the reported 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms,[50] and 2,851 clusters for 171,045 natural products.[51] A total of 42,670−44,115 compounds extracted from the 8,534−8,823 families (5 per family) that contain no known inhibitor were used as the putative noninhibitors.

(35) Li, J.; Gramatica, P. Classification and Virtual Screening of Androgen Receptor Antagonists. *J. Chem. Inf. Model.* **2010**, *50*, 861–74.

(36) Derksen, S.; Rau, O.; Schneider, P.; Schubert-Zsilavecz, M.; Schneider, G. Virtual screening for PPAR modulators using a probabilistic neural network. *ChemMedChem* **2006**, *1* (12), 1346–50.

(37) Irwin, J. J.; Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–82.

(38) Noble, M. E.; Endicott, J. A.; Johnson, L. N. Protein kinase inhibitors: insights into drug design from structure. *Science* **2004**, *303* (5665), 1800–5.

(39) Vema, A.; Panigrahi, S. K.; Rambabu, G.; Gopalakrishnan, B.; Sarma, J. A.; Desiraju, G. R. Design of EGFR kinase inhibitors: a ligand-based approach and its confirmation with structure-based studies. *Bioorg. Med. Chem.* **2003**, *11* (21), 4643–53.

(40) Yu, H.; Wang, Z.; Zhang, L.; Zhang, J.; Huang, Q. Pharmacophore modeling and in silico screening for new KDR kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17* (8), 2126–33.

(41) Matsuno, K.; Ichimura, M.; Nakajima, T.; Tahara, K.; Fujiwara, S.; Kase, H.; Ushiki, J.; Giese, N. A.; Pandey, A.; Scarborough, R. M.; Lokker, N. A.; Yu, J. C.; Irie, J.; Tsukuda, E.; Ide, S.; Oda, S.; Nomoto, Y. Potent and selective inhibitors of platelet-derived growth factor receptor phosphorylation. 1. Synthesis, structure-activity relationship, and biological effects of a new class of quinazoline derivatives. *J. Med. Chem.* **2002**, *45* (14), 3057–66.

(42) Thompson, A. M.; Connolly, C. J.; Hamby, J. M.; Boushelle, S.; Hartl, B. G.; Amar, A. M.; Kraker, A. J.; Driscoll, D. L.; Steinkampf, R. W.; Patmore, S. J.; Vincent, P. W.; Roberts, B. J.; Elliott, W. L.; Klohs, W.; Leopold, W. R.; Showalter, H. D.; Denny, W. A. 3-(3,5-Dimethoxyphenyl)-1,6-naphthyridine-2,7-diamines and related 2-urea derivatives are potent and selective inhibitors of the FGF receptor-1 tyrosine kinase. *J. Med. Chem.* **2000**, *43* (22), 4200–11.

(43) Dalgarno, D.; Stehle, T.; Narula, S.; Schelling, P.; van Schravendijk, M. R.; Adams, S.; Andrade, L.; Keats, J.; Ram, M.; Jin, L.; Grossman, T.; MacNeil, I.; Metcalf, C., 3rd; Shakespeare, W.; Wang, Y.; Keenan, T.; Sundaramoorthi, R.; Bohacek, R.; Weigele, M.; Sawyer, T. Structural basis of Src tyrosine kinase inhibition with a new class of potent and selective trisubstituted purine-based compounds. *Chem. Biol. Drug Des.* **2006**, *67* (1), 46–57.

(44) Abbott, L.; Betschmann, P.; Burchat, A.; Calderwood, D. J.; Davis, H.; Hrnciar, P.; Hirst, G. C.; Li, B.; Morytko, M.; Mullen, K.; Yang, B. Discovery of thienopyridines as Src-family selective Lck inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17* (5), 1167–71.

(45) Showalter, H. D.; Sercel, A. D.; Leja, B. M.; Wolfangel, C. D.; Ambroso, L. A.; Elliott, W. L.; Fry, D. W.; Kraker, A. J.; Howard, C. T.; Lu, G. H.; Moore, C. W.; Nelson, J. M.; Roberts, B. J.; Vincent, P. W.; Denny, W. A.; Thompson, A. M. Tyrosine kinase inhibitors. 6. Structure-activity relationships among N- and 3-substituted 2,2′-diselenobis(1H-indoles) for inhibition of protein tyrosine kinases and comparative in vitro and in vivo studies against selected sulfur congeners. *J. Med. Chem.* **1997**, *40* (4), 413–26.

(46) Asano, T.; Yoshikawa, T.; Usui, T.; Yamamoto, H.; Yamamoto, Y.; Uehara, Y.; Nakamura, H. Benzamides and benzamidines as specific inhibitors of epidermal growth factor receptor and v-Src protein tyros ine kinases. *Bioorg. Med. Chem.* **2004**, *12* (13), 3529–42.

(47) Caballero, J.; Fernandez, M.; Saavedra, M.; Gonzalez-Nilo, F. D. 2D Autocorrelation, CoMFA, and CoMSIA modeling of protein tyrosine kinases' inhibition by substituted pyrido[2,3-d]pyrimidine derivatives. *Bioorg. Med. Chem.* **2008**, *16* (2), 810–21.

(48) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35* (Database issue), D198–201.

(49) Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem* **2001**, *3* (2), 157–66.

(50) Fink, T.; Reymond, J.-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**, *47* (2), 342–53.

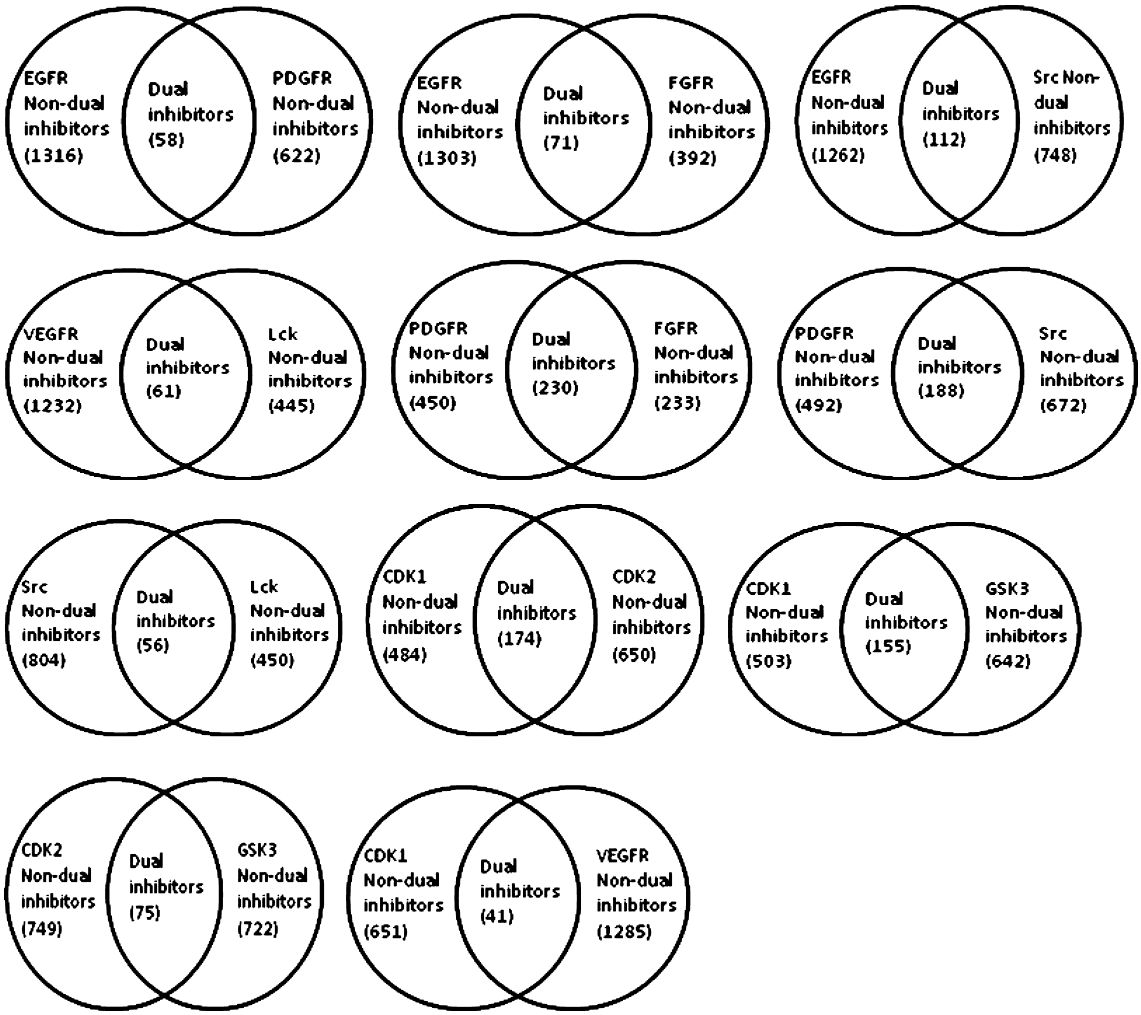
EGFR Non-dual inhibitors (1316)
Dual inhibitors (58)
PDGFR Non-dual inhibitors (622)
EGFR Non-dual inhibitors (1303)
Dual inhibitors (71)
FGFR Non-dual inhibitors (392)
EGFR Non-dual inhibitors (1262)
Dual inhibitors (112)
Src Non-dual inhibitors (748)
VEGFR Non-dual inhibitors (1232)
Dual inhibitors (61)
Lck Non-dual inhibitors (445)
PDGFR Non-dual inhibitors (450)
Dual inhibitors (230)
FGFR Non-dual inhibitors (233)
PDGFR Non-dual inhibitors (492)
Dual inhibitors (188)
Src Non-dual inhibitors (672)
Src Non-dual inhibitors (804)
Dual inhibitors (56)
Lck Non-dual inhibitors (450)
CDK1 Non-dual inhibitors (484)
Dual inhibitors (174)
CDK2 Non-dual inhibitors (650)
CDK1 Non-dual inhibitors (503)
Dual inhibitors (155)
GSK3 Non-dual inhibitors (642)
CDK2 Non-dual inhibitors (749)
Dual inhibitors (75)
GSK3 Non-dual inhibitors (722)
CDK1 Non-dual inhibitors (651)
Dual inhibitors (41)
VEGFR Non-dual inhibitors (1285)

Figure 1. The Venn graph of the collected dual-inhibitors, the 11 evaluated kinase pairs and non-dual-inhibitors of the 9 evaluated kinases.

The collected non-dual- and dual-inhibitors of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3 are distributed in 431, 456, 246, 170, 284, 192, 255, 301, and 295 families respectively, which is consistent with reported 191 unique scaffolds (154 clusters and 43 singletons) for 565 kinase inhibitors.[17] 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 of families vs 8,534−8,823 families contained in PubChem and MDDR at present) is expected to be no more than ∼999/8500, which is <13%. Therefore, putative noninhibitor training data set can be generated by extracting a few representative compounds from each of the families that contain no known inhibitor, with a maximum possible "wrong" classification rate of <13% even in the extreme and unlikely cases that all of the undiscovered inhibitors are misplaced into the noninhibitor 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 SVM.[52] It is noted that 40%−62.2% of the dual-inhibitor families contain no non-dual-inhibitor of the same kinase pair, whose representative compounds were included in the inactive training data sets as dual-inhibitors and 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 C-SVM.

Molecular descriptors quantitatively represent structural and physicochemical features of molecules, and have been extensively used in deriving structure−activity relation-

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

(52) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. 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), 193–200.

Table 2. Molecular Descriptors Used in This Work

descriptor class	no. of descriptors in class	descriptors
simple molecular properties[82]	18	no. of C, N, O, P, S, no. of total atoms, no. of rings, no. of bonds, no. of non-H bonds, MW, no. of rotatable bonds, no. of H-bond donors, no. of H-bond acceptors, no. of 5-member aromatic rings, no. of 6-member aromatic rings, no. of N heterocyclic rings, no. of O heterocyclic rings, no. of S heterocyclic rings
chemical properties[83]	3	Sanderson electronegativity, molecular polarizability, aLogp
molecular connectivity and shape[82,84]	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*, 0−2nd valence connectivity index, 0−2nd order delta chi index, Pogliani index, 0−2nd solvation connectivity index, 1−3rd order Kier shape index, 1−3rd order kappa alpha shape index, Kier molecular flexibility index, topological radius, graph−theoretical shape coefficient, eccentricity, centralization, Logp from connectivity
electrotopological state[82,85]	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

ships,[53] quantitative structure−activity relationships[54] and VS tools.[13,17,21] We used 98 1D and 2D descriptors (Table 2) derived from our software,[55] which include 18 descriptors in the class of simple molecular properties, 3 descriptors in the class of chemical properties, 35 descriptors in the class of molecular connectivity and shape, and 42 descriptors in the class of electrotopological state.

Support Vector Machines Methodology. SVM training and its application is schematically illustrated in Figure 2. SVM is based on the structural risk minimization principle of statistical learning theory.[56] It consistently shows outstanding classification performance, is less penalized by sample redundancy, has lower risk for overfitting, is capable of accommodating large and structurally diverse training and testing data sets, and is fast in performing classification tasks.[57,58] However, like all machine learning methods, the performance of SVM is critically dependent on the diversity of training data sets. Our earlier study showed that the yields of SVM trained by sparse data sets can be reduced by 10%−20% if the training data set is reduced by 50%, and the reduction is in a compound diversity dependent manner with higher reduction percentage for higher diversity compounds.[13] Because of the limited knowledge of known inhibitors for many kinase targets, sufficiently good SVM VS tools may not be readily developed for these targets. Nonetheless, SVM VS tools with comparable performances or partially improved performances in certain aspects (e.g., reduced false-hit rates at comparable inhibitor yield) are useful to complement other VS tools. The detailed mathematical algorithms of SVM are described in our earlier publications[13,21,55,59] as well as in the established literature.[56−58] Readers are referred to this literature. Our SVM VS models were developed by using a hard margin $c = 100{,}000$, and their σ values are in the range of 0.1−2. In terms of the numbers of true positives TP (true inhibitors), true negatives TN (true noninhibitors), false positives FP (false inhibitors), and false negatives FN (false noninhibitors), the yield and false-hit rate are given by TP/(TP + FN) and FP/(TP + FP) respectively.

Results and Discussion

Dual-Inhibitors and Non-Dual-Inhibitors of the Studied Kinase Pairs. As shown in Table 1, the numbers of dual-inhibitors and non-dual-inhibitors of the kinase pairs are 58, 1,316 and 622 for EGFR-PDGFR, 71, 1,303 and 392 for

[53] Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112* (12), 1249–1254.

[54] Ijjaali, I.; Petitet, F.; Dubus, E.; Barberan, O.; Michel, A. Assessing potency of c-Jun N-terminal kinase 3 (JNK3) inhibitors using 2D molecular descriptors and binary QSAR methodology. *Bioorg. Med. Chem.* **2007**, *15* (12), 4256–64.

[55] Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1497–505.

[56] Vapnik, V. N. *The nature of statistical learning theory*; Springer: New York, 1995.

[57] Pochet, N.; De Smet, F.; Suykens, J. A.; De Moor, B. L. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **2004**, *20*, 3185–95.

[58] Li, F.; Yang, Y. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics* **2005**, *21*, 3741–3747.

[59] Ung, C. Y.; Li, H.; Yap, C. W.; Chen, Y. Z. In silico prediction of pregnane X receptor activators by machine learning approaches. *Mol. Pharmacol.* **2007**, *71* (1), 158–68.

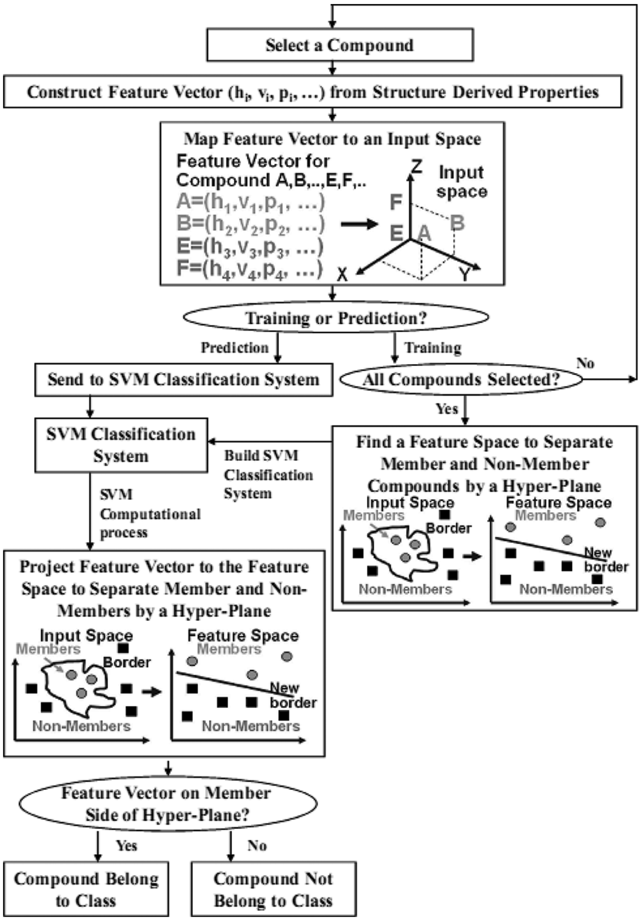
Select a Compound
Construct Feature Vector (h_i, v_i, p_i, ...) from Structure Derived Properties
Map Feature Vector to an Input Space
Feature Vector for Compound A,B,..,E,F,..
A=(h1,v1,p1, ...)
B=(h2,v2,p2, ...)
E=(h3,v3,p3, ...)
F=(h4,v4,p4, ...)
Z Input space
F E A B X Y
Training or Prediction?
Prediction
Training
Send to SVM Classification System
All Compounds Selected?
No
Yes
SVM Classification System
Build SVM Classification System
SVM Computational process
Find a Feature Space to Separate Member and Non-Member Compounds by a Hyper-Plane
Input Space
Feature Space
Members
Border
New border
Non-Members
Project Feature Vector to the Feature Space to Separate Member and Non-Members by a Hyper-Plane
Input Space
Feature Space
Members
Border
New border
Non-Members
Feature Vector on Member Side of Hyper-Plane?
Yes
No
Compound Belong to Class
Compound Not Belong to Class

Figure 2. Schematic diagram illustrating the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally derived properties (molecular descriptors) by using support vector machines. A, B, E, F and (h_j, p_j, v_j, . . .) represent such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

EGFR-FGFR, 112, 1,262 and 748 for EGFR-Src, 61, 1,232 and 445 for VEGFR-Lck, 230, 450, and 233 for PDGFR-FGFR, 188, 492, and 672 for PDGFR-Src, 56, 804, and 450 for Src-Lck, 174, 484, and 650 for CDK1-CDK2, 155, 503, and 642 for CDK1-GSK3, 75, 749, and 722 for CDK2-GSK3, and 41, 651, and 1285 for CDK1-VEGFR respectively. The dual-inhibitors and non-dual-inhibitors are distributed in 17−84 and 90−427 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 data sets are 62.1% for EGFR-PDGFR, 57.9% for EGFR-FGFR, 58.9% for EGFR-Src, 52.5% for VEGFR-Lck, 60.9% for PDGFR-FGFR, 62.2% for PDGFR-Src, 58.9% for Src-Lck, 69.5% for CDK1-CDK2, 68.4% for CDK1-GSK3, 58.7% for CDK2-GSK3, and 82.9% for CDK1-VEGFR respectively. Therefore, these dual-inhibitors have a substantial degree of novelty against non-dual-inhibitors. Moreover, 0.0%−98.6% of the dual-inhibitors of the kinase pairs are inhibitor of at least one of the other 7 kinases, but only up to 5.2% of the dual-inhibitors are inhibitors of at least 3 of the other 7 kinases. Hence, most of these dual-inhibitors are nonubiquitous inhibitors and show some degree of kinase selectivity even though the majority of them target more than 2 kinases.

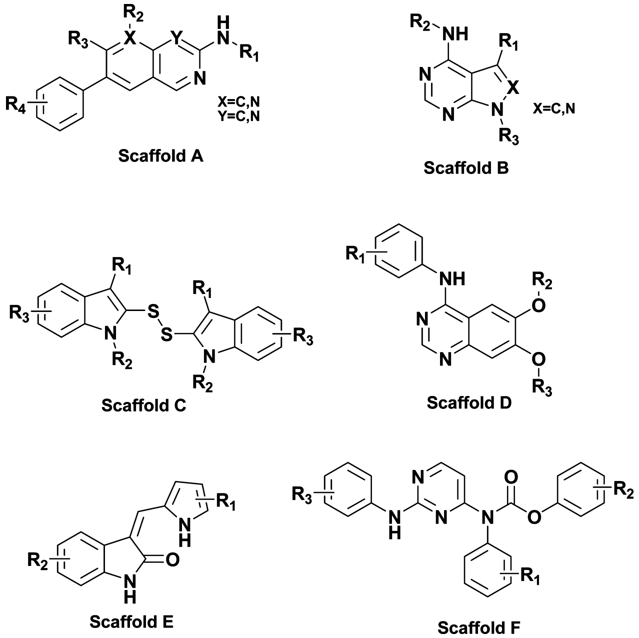

Figure 3. Top 6 scaffolds contained in higher percentages of the dual-inhibitors of the studied intra-PTK group kinase pairs.

Some distinguishing features of dual-inhibitors may be probed by evaluating the top 6 scaffolds contained in higher percentages of the dual-inhibitors of the studied intra-PTK group kinase pairs, which are shown in Figure 3. Table 3 shows the distribution of these scaffolds in the dual-inhibitors and non-dual-inhibitors of the studied intra-PTK group kinase pairs. Scaffold A is contained in 63.8% of EGFR-PDGFR, 76.1% of PDGFR-Src, 33.9% of EGFR-Src, 54.9% of EGFR-FGFR and 27.8% of VEGFR-Lck dual-inhibitors respectively; scaffold B is contained in 57.1% of Src-Lck, 29.5% of VEGFR-Lck and 25.9% of EGFR-Src dual-inhibitors respectively. Scaffold A and scaffold B appear to be the backbone of majority of dual-inhibitors of the studied kinase pairs. Scaffold C is mainly contained in 19.6% of EGFR-Src dual-inhibitors. Scaffold D is mainly contained in 32.4% in EGFR-FGFR and 4.5% in EGFR-Src dual-inhibitors. Scaffold E is contained in 17.8% of PDGFR-FGFR, 8.6% of EGFR-PDGFR, 7.0% of EGFR-FGFR and 6.9% of PDGFR-Src dual-inhibitors. Scaffold F is contained in 37.5% of Src-Lck and 34.4% of VEGFR-Lck dual-inhibitors. These scaffolds are also contained, mostly at significantly lower percentage levels, in the non-dual-inhibitors of at least one of the kinases of the respective kinase pairs. Therefore, some specific variations of side-chain groups of these scaffolds appear to be sufficient to convert some dual-inhibitors into non-dual-inhibitors, which suggests that physicochemical properties as well as structural features are important for distinguishing dual- and non-dual-inhibitors.

Table 3. Distribution of Top 6 Scaffolds in Dual-Inhibitors of 7 Intra-PTK Group Kinase Combinations of EGFR, VEGFR, PDGFR, FGFR, Src and Lck, and Non-Dual-Inhibitors of the Constituent Kinases

		percentage of inhibitors containing scaffold					
kinase pair	data sets	A	B	C	D	E	F
EGFR-PDGFR	dual-inhibitors	63.8 (37/58)	0 (0/58)	0 (0/58)	1.7 (1/58)	8.6 (5/58)	0 (0/58)
	EGFR non-dual-inhibitors	0.2 (3/1316)	6.3 (83/1316)	1.2 (16/1316)	7.7 (101/1316)	0 (0/1316)	0 (0/1316)
	PDGFR non-dual-inhibitors	20.3 (126/622)	0 (0/622)	0 (0/622)	0 (0/622)	7.1 (44/622)	0 (0/622)
EGFR-FGFR	dual-inhibitors	54.9 (39/71)	0 (0/71)	0 (0/71)	32.4 (23/71)	7.0 (5/71)	0 (0/71)
	EGFR non-dual-inhibitors	0.1 (1/1303)	6.4 (83/1303)	1.2 (16/1303)	6.1 (79/1303)	0 (0/1303)	0 (0/1303)
	FGFR non-dual-inhibitors	25.5 (100/392)	0 (0/392)	0 (0/392)	2.3 (9/392)	10.0 (39/392)	0.3 (1/392)
EGFR-Src	dual-inhibitors	33.9 (38/112)	25.9 (29/112)	19.6 (22/112)	4.5 (5/112)	2.7 (3/112)	0 (0/112)
	EGFR non-dual-inhibitors	0.2 (2/1262)	4.3 (54/1262)	1.6 (20/1262)	7.7 (97/1262)	0.2 (2/1262)	0 (0/1262)
	Src non-dual-inhibitors	18.2 (136/748)	10.4 (78/748)	0.8 (6/748)	5.0 (37/748)	1.60 (12/748)	2.8 (21/748)
VEGFR-Lck	dual-inhibitors	27.9 (17/61)	29.5 (18/61)	0 (0/61)	0 (0/61)	0 (0/61)	34.4 (21/61)
	VEGFR non-dual-inhibitors	0.7 (8/1232)	0.8 (10/1232)	0 (0/1232)	5.4 (66/1232)	4.7 (58/1232)	0 (0/1232)
	Lck non-dual-inhibitors	5.6 (25/445)	10.3 (46/445)	0 (0/445)	1.6 (7/445)	0 (0/445)	1.6 (7/445)
PDGFR-FGFR	dual-inhibitors	67.4 (155/230)	0 (0/230)	0 (0/230)	0 (0/230)	17.8 (41/230)	0 (0/230)
	PDGFR non-dual-inhibitors	1.8 (8/450)	0 (0/450)	0 (0/450)	0.2 (1/450)	1.8 (8/450)	0 (0/450)
	FGFR non-dual-inhibitors	11.2 (26/233)	0 (0/233)	0 (0/233)	13.7 (32/233)	1.3 (3/233)	0.4 (1/233)
PDGFR-Src	dual-inhibitors	76.1 (143/188)	0 (0/188)	0 (0/188)	0 (0/188)	6.9 (13/188)	0 (0/188)
	PDGFR non-dual-inhibitors	2.9 (14/492)	0 (0/492)	0 (0/492)	0.2 (1/492)	7.3 (36/492)	0 (0/492)
	Src non-dual-inhibitors	3.7 (25/672)	15.9 (107/672)	1.9 (13/672)	6.3 (42/672)	0.3 (2/672)	3.1 (21/672)
Src-Lck	dual-inhibitors	0 (0/56)	57.1 (32/56)	0 (0/56)	1.8 (1/56)	1.8 (1/56)	37.5 (21/56)
	Src non-dual-inhibitors	21.6 (174/804)	9.3 (75/804)	1.6 (13/804)	5.1 (41/804)	1.9 (15/804)	0 (0/804)
	Lck non-dual-inhibitors	5.9 (26/450)	7.8 (35/450)	0 (0/450)	1.3 (6/450)	0 (0/450)	2 (9/450)

Virtual Screening Performance of Combinatorial SVM in Searching Kinase Dual-Inhibitors from Large Libraries. The VS performance of C-SVM in identifying dual-inhibitors of the 11 kinase pairs is summarized in Table 4 and further shown in Figure 4. The parameters of the developed SVM classification models for the evaluated kinases are in the range of $\sigma = 0.5-0.8$. The dual-inhibitor yields are 27.6% for EGFR-PDGFR, 40.9% for EGFR-FGFR, 26.8% for EGFR-Src, 52.6% for VEGFR-Lck, 33.9% for PDGFR-FGFR, 38.3% for PDGFR-Src, 48.2% for Src-Lck, 52.3% for CDK1-CDK2, 49.0% for CDK1-GSK3, 57.3% for CDK2-GSK3, and 12.2% for CDK1-VEGFR respectively. The yields for the intra-PTK group and intra-CMGC group kinase pairs are comparable to the expected 25%−49% yields of combinations of good VS tools with individual yields of 50%−70%. Therefore, C-SVM show reasonably good capability in identifying multitarget agents for kinase pairs within a protein kinase group without requiring explicit knowledge of multitarget agents. However, the yield for the inter-PTK-CMGC kinase group CDK1-VEGFR kinase pair is only 12.2%, which is significantly lower than those for the intra-PTK group and intra-CMGC group kinase pairs. Structural analysis of the inhibitors of CDK1 and VEGFR binding sites has revealed that inhibitors generally make extensive favorable van der Waals contacts and several hydrogen bonds with Lys33, Leu83 and Asp86 at the hinge region of CDK1, and with Cys919, Asn923, Cys1045 and Asp1046 at the hinge region of VEGFR respectively, relatively small structural changes may easily reduce the optimal fit to the binding site, and some dual-inhibitors are able to bind to both kinases because of their structural flexibility to tolerate the different binding site geometry and to form alternative hydrogen bonds.[60] In some cases, dual selectivity of inhibitors of inter-kinase-group kinase pairs may require structural flexibility to fit in a hydrophobic pocket conserved in both kinase classes.[61] Such special structural features in dual-inhibitors of interkinase-group kinase pairs are not necessarily needed and thus may not be found in non-dual-inhibitors of individual kinases used in our training data sets, which could likely be an important reason for the reduced yield of C-SVM in identifying CDK1-VEGFR dual-inhibitors. The smaller number of known CDK1-VEGFR dual-inhibitors may also affect the accurate assessment of VS outcome.

Target selectivity was tested by using C-SVM to screen the 233−1,316 non-dual-inhibitors of the 11 kinase pairs, which misidentified 9.2% and 14.3% of the non-dual-inhibitors of the kinase pair as dual-inhibitors for EGFR-PDGFR, 10.1% and 8.7% for EGFR-FGFR, 12.9% and 11.1% for EGFR-Src, 6.6% and 29.2% for VEGFR-Lck, 15.6% and 22.3% for PDGFR-FGFR, 25.8% and 11.6% for

[60] Kuo, G. H.; Wang, A.; Emanuel, S.; Deangelis, A.; Zhang, R.; Connolly, P. J.; Murray, W. V.; Gruninger, R. H.; Sechler, J.; Fuentes-Pesquera, A.; Johnson, D.; Middleton, S. A.; Jolliffe, L.; Chen, X. Synthesis and discovery of pyrazine-pyridine biheteroaryl as a novel series of potent vascular endothelial growth factor receptor-2 inhibitors. *J. Med. Chem.* **2005**, *48* (6), 1886–900.

[61] Apsel, B.; Blair, J. A.; Gonzalez, B.; Nazif, T. M.; Feldman, M. E.; Aizenstein, B.; Hoffman, R.; Williams, R. L.; Shokat, K. M.; Knight, Z. A. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat. Chem. Biol.* **2008**, *4* (11), 691–9.

Table 4. Virtual Screening Performance of Combinatorial SVM for Identifying Dual-Inhibitors of 11 Combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3

	virtual screening performance				
	dual inhibitors		non-dual-inhibitors of the same kinase pair		false hit rate
kinase	yield, %	no. (%) of identified true hits outside the common training active families of both kinases	false hit rate for inhibitors of kinase A, %	false hit rate for inhibitors of kinase B, %	inhibitors of other 7 kinases, %
EGFR-PDGFR	27.60	9 (15.5)	9.20	14.30	1.88
EGFR-FGFR	40.90	6 (8.5)	10.10	8.70	1.06
EGFR-Src	26.80	13 (11.6)	12.90	11.10	1.49
VEGFR-Lck	52.60	8 (13.1)	6.60	29.20	2.80
PDGFR-FGFR	33.90	35 (15.2)	15.60	22.30	0.98
PDGFR-Src	38.30	30 (16.0)	25.80	11.60	1.81
Src-Lck	48.20	9 (16.1)	15.80	18.70	0.98
CDK1-CDK2	52.30	57 (32.8)	39.20	48.10	3.39
CDK1-GSK3	49.00	41 (26.5)	38.40	37.40	4.30
CDK2-GSK3	57.30	24 (32.0)	36.80	37.70	2.99
CDK1-VEGFR	12.20	0 (0.0)	14.00	3.70	4.77

	virtual screening performance			
	virtual hit rate, % (no. of virtual hits)			
kinase	MDDR compounds similar to dual-inhibitors	all 168 thousand MDDR compounds	13.56 million PubChem compounds	1.02 million Zinc clean-leads data set
EGFR-PDGFR	1.5 (57)	0.10 (175)	0.031 (4155)	0.025 (257)
EGFR-FGFR	6.5 (65)	0.07 (126)	0.016 (2200)	0.004 (36)
EGFR-Src	2.13 (24)	0.096 (162)	0.033 (4471)	0.007 (76)
VEGFR-Lck	5.1 (21)	0.10 (170)	0.036 (4817)	0.011 (113)
PDGFR-FGFR	1.4 (51)	0.057 (95)	0.013 (1746)	0.0008 (8)
PDGFR-Src	2.9 (84)	0.104 (175)	0.021 (2799)	0.001 (14)
Src-Lck	9.4 (26)	0.078 (131)	0.020 (2674)	0.002 (25)
CDK1-CDK2	0.34 (9)	0.075 (126)	0.022 (2953)	0.014 (139)
CDK1-GSK3	0.30 (10)	0.028 (47)	0.016 (2218)	0.016 (159)
CDK2-GSK3	0.43 (7)	0.085 (142)	0.021 (2901)	0.020 (203)
CDK1-VEGFR	0.0 (0)	0.007 (12)	0.023 (3113)	0.002 (19)

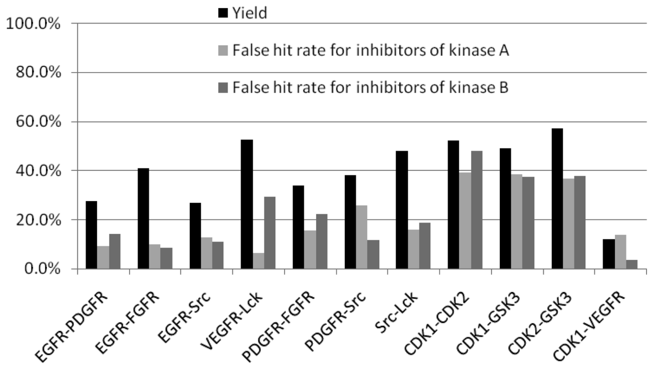

Figure 4. The VS performance of C-SVM in identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3.

PDGFR-Src, 15.8% and 18.7% for Src-Lck, 39.2% and 48.1% for CDK1-CDK2, 38.4% and 37.4% for CDK1-GSK3, 36.8% and 37.7% for CDK2-GSK3, and 14.0% and 3.7% for CDK1-VEGFR respectively. Therefore, C-SVM are reasonably selective in distinguishing dual-inhibitors from non-dual-inhibitors. There are two possible reasons for the misidentification of a substantial percentage of non-dual-inhibitors as dual-inhibitors. First, SVMs were trained by non-dual-inhibitors only, which may not fully distinguish dual- and non-dual-inhibitors. Second, some of the misidentified non-dual-inhibitors are probably true dual-inhibitors not yet experimentally tested for multitarget 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 C-SVM to screen the 3,971−5,180 inhibitors of the other 7 kinases not included in a particular kinase pair. We found that 1.88% of these inhibitors were misidentified as dual-inhibitors for EGFR-PDGFR, 1.06% for EGFR-FGFR, 1.49% for EGFR-Src, 2.80% for VEGFR-Lck, 0.98% for PDGFR-FGFR, 1.81% for PDGFR-Src, 0.98% for Src-Lck, 3.39% for CDK1-CDK2, 4.30% for CDK1-GSK3, 2.99% for CDK2-GSK3, and 4.77% for CDK1-VEGFR respectively. These showed that C-SVM are fairly selective in separating inhibitors of specific kinase pair from those of other kinases.

Virtual-hit rates and false-hit rates of C-SVM in screening compounds that resemble the structural and physicochemical properties of the training data sets were evaluated by using 276−3,614 MDDR compounds similar to a dual-inhibitor of each kinase pair. Similarity was defined by Tanimoto similarity coefficient ≥0.9 between a MDDR compound and its closest dual-inhibitor.[13] C-SVM identified 57 virtual-hits from 3,806 MDDR similarity compounds (virtual-hit rate 1,5%) for EGFR-PDGFR, 65 from 1,001 MDDR compounds (6.5%) for EGFR-FGFR, 24 from 1,127 MDDR compounds (2.1%) for EGFR-Src, 21 from 413 MDDR compounds (5.1%) for VEGFR-Lck, 51 from 3,614 MDDR compounds (1.4%) for PDGFR-FGFR, 84 from 2,893 MDDR compounds (2.9%) for PDGFR-Src, 26 from 276 MDDR compounds (9.4%) for Src-Lck, 9 from 2,629 MDDR compounds (0.34%) for CDK1-CDK2, 10 from 3,279 MDDR compounds (0.30%) for CDK1-GSK3, 7 from 1,617 MDDR compounds (0.43%) for CDK2-GSK3, and 0 from 505 MDDR compounds (0.0%) for CDK1-VEGFR 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 175 and 0.1% for EGFR-PDGFR, 126 and 0.07% for EGFR-FGFR, 162 and 0.096% for EGFR-Src, 170 and 0.1% for VEGFR-Lck, 95 and 0.057% for PDGFR-FGFR, 175 and 0.104% for PDGFR-Src, and 131 and 0.078% for Src-Lck, 126 and 0.075% for CDK1-CDK2, 47 and 0.028% for CDK1-GSK3, 142 and 0.085% for CDK2-GSK3 and 12 and 0.007% for CDK1-VEGFR respectively. The numbers of virtual-hits and virtual-hit rates in screening 13.56 M PubChem compounds are 4,155 and 0.031% for EGFR-PDGFR, 2,200 and 0.015% for EGFR-FGFR, 4,471 and 0.033% for EGFR-Src, 4,817 and 0.036% for VEGFR-Lck, 1,746 and 0.013% for PDGFR-FGFR, 2,799 and 0.021% for PDGFR-Src, 2,674 and 0.02% for Src-Lck, 2,953 and 0.022% for CDK1-CDK2, 2,218 and 0.016% for CDK1-GSK3, 2,901 and 0.021% for CDK2-GSK3, and 3,113 and 0.023% for CDK1-VEGFR respectively.

Substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, tyrosine-specific protein kinase inhibitors, and signal transduction inhibitors (Table 6, details in next section). As some of these virtual-hits may be true dual-inhibitors, the actual number of true false-hits may be smaller than the total number of virtual-hits for each kinase pair. Hence, the false-hit rates of our combinatorial SVMs are at most equal to and likely less than the virtual-hit rates. Hence the false-hit rates are ≤1.4%−9.4% in screening 276−3,614 MDDR similarity compounds, ≤0.057%−0.104% in screening 168 thousand MDDR compounds, and ≤0.013%−0.036% 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 SVMs,[12,13] 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.[62]

Comparison of the Performance of Combinatorial SVM with Other Virtual Screening Methods. The VS performance of C-SVM was further compared with Surflex-Dock,[63] DOCK 3.5.54 at the DOCK Blaster server,[34] kNN,[35] and PNN[36] by using the common testing data sets composed of 41−230 dual-inhibitors of the 11 evaluated kinase pairs (set-1), 3,971−5,180 non-dual-inhibitors of the 9 evaluated kinases (set-2), and 1.02 million Zinc clean-leads data set (Zinc-CLD)[37] (set-3) respectively. Surflex-Dock VS study was conducted on set-1, set-2, and a Zinc-CLD subset (set-

[62] Ma, X. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Chen, Y. Z. Comparative Analysis of Machine Learning Methods in Ligand Based Virtual Screening of Large Compound Libraries. *Comb. Chem. High Throughput Screen.* **2009**, *12* (4), 344–57.

[63] Jain, A. N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21* (5), 281–306.

4) generated by clustering the 1.02 million compounds in set-3 into 10,000 clustered followed by the selection of one representative compound from each cluster. Clustering of the set-3 compounds was conducted by using the K-means algorithm with each compound represented by the 98 1D and 2D descriptors derived from our software.[55] The compound closest to the centroid of each cluster was selected as the representative compound of that cluster. Surflex-Dock and DOCK Blaster VS studies were conducted against the protein crystal structures typically used in DOCK Blaster VS studies.[34] Specifically, the PDB entry for EGFR, FGFR, c-Src, VEGFR, CDK2, Lck, and GSK3 are 3BEL, 3C4F, 1YOL, 1Y6B, 2A4L, 2OG8, and 1Q5K respectively.[34] Moreover, a modeled 3D structure of PDGFR in the well-established molecular docking benchmarking sets[64] was used for PDGFR. CDK1 was not evaluated because we were unable to find a published experimental or modeled 3D structure.

In Surflex-Dock and DOCK Blaster VS studies, the dual-inhibitor yield was estimated based on the screening results of set-1 and set-2 compounds, which is the percentage of the known dual-inhibitors made to the top-50% of the successfully docked set-1 and set-2 compounds for every kinase of a kinase pair, the false-hit rate for misidentifying inhibitors of other 7 kinases as dual-inhibitors of a kinase pair is the percentage of these inhibitors made to the top-50% of the successfully docked set-1 and set-2 compounds for every kinase of that kinase pair, and the virtual-hit rate for the Zinc-CLD compounds is the percentage of these compounds made to the top-2% of the successfully docked set-3 or set-4 compounds for every kinase of that kinase pair. The kNN and PNN methods and software used in this study were described in our previous studies[55,59] and summarized in the Supporting Information. The training data sets of kNN and PNN and the methods for estimating the yield and virtual hit rate are the same as those of SVM. The parameters of the developed kNN and PNN classification models for the evaluated kinases are in the ranges of $k = 1$ or 3, and $\delta = 0.003-0.11$ respectively. The CPU time is ~0.12, ~8, and ~5.5 h per kinase target of SVM, kNN, and PNN models in screening the 1.02 million Zinc clean-leads data set respectively. The classification speed of SVM is faster than that of kNN and PNN due to the fact that SVM typically uses 0.007−0.017% of the training data set as support vectors for classification, whereas kNN and PNN use the whole training data set. It took ~35 h for using Surflex-Dock to screen the 10K subset of the Zinc ckean-leads data set, and ~2 weeks for getting the docking results from the DOCK Blaster server for screening the whole Zinc clean-leads data set per kinase target.

Table 5 and Figure 5 shows the comparison of the performance of C-SVM with the other four VS methods for identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3 from the four common testing data sets (set-1, set-2, set-3 and set-4). Overall, the yields of all VS methods are comparable, mostly in the ranges of 21.3%−57.3% for the intra-PTK and intra-CMGC group kinase pairs and 12.2%−19.5% for the inter-PTK-CMGC group kinase pair. C-SVM, kNN and PNN also produced comparable false hit-rates, at 0.98%−6.05%, for misidentifying inhibitors of other 7 kinases as dual-inhibitors of the evaluated kinase pairs, with SVM showing slightly lower false hit-rates for the majority of the evaluated kinase pairs.

For the 8 kinase pairs with available 3D structure, Surflex-Dock and DOCK Blaster produced higher false hit-rates than other three evaluated VS methods in misidentifying inhibitors of other 7 kinases as dual-inhibitors. These false-hit rates may be significantly reduced by adjusting the docking cutoff values for individual kinases, e.g. from top-50% to top-10%, which may however lead to significantly reduced yields. High false-positive rates has been a common issue in structure-based VS, and the false-positives in kinase docking studies arise partly from the inability to favorably score certain key hydrogen-binding interactions required for kinase binding and to discriminate conformational artifacts of docked ligands.[65] False-hit rates can be significantly reduced by such strategies as the incorporation of the reported kinase binding features into docking constrains,[65] consensus scoring using multiple ligand information and maximum common binding modes for multiple kinases,[66] and combining docking with pharmacophore filtering.[67]

In screening the Zinc-CLD whole-set and subset, C-SVM produced substantially lower virtual-hit rates (0.008%−0.025% and 0.00%−0.05% respectively) than those (0.009%−0.348% and 0.03%−0.37% respectively) of the other four VS methods in identifying the Zinc-CLD compounds as virtual dual-inhibitors of the evaluated kinase pairs. It is noted that the virtual-hit rates in screening the Zinc-CLD subset are comparable to those of the Zinc-CLD whole-set, suggesting that virtual screening performance on Zinc-CLD may be probed by using the Zinc-CLD subset. The numbers of Zinc-CLD compounds identified as virtual-hits by C-SVM are in the range of 8−203 and 0−5 in screening the Zinc-CLD whole-set and subset respectively, compared to those of 1439−3963, 96−1406, and 332−2830 by DOCK Blaster, kNN, and PNN in screening the Zinc-CLD whole-set, and those of 14−25, 5−31, 2−15, and 3−37 by Surflex-Dock, DOCK Blaster, kNN, and PNN in screening the Zinc-CLD subset, respectively.

The numbers of undiscovered dual-inhibitors of the evaluated kinase pairs in the Zinc-CLD are unknown. It is

[64] Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–801.

[65] Perola, E. Minimizing false positives in kinase virtual screens. *Proteins* **2006**, *64* (2), 422–35.

[66] Renner, S.; Derksen, S.; Radestock, S.; Morchen, F. Maximum common binding modes (MCBM): consensus docking scoring using multiple ligand information and interaction fingerprints. *J. Chem. Inf. Model.* **2008**, *48* (2), 319–32.

[67] Peach, M. L.; Nicklaus, M. C. Combining docking with pharmacophore filtering for improved virtual screening. *J. Cheminf.* **2009**, *1* (1), 6.

Table 5. Comparison of the Performance of Combinatorial SVM with Other Virtual Screening Methods for Identifying Dual-Inhibitors of 11 Combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3[a]

	virtual screening performance for kinase pair, %										
methods	EGFR-PDGFR	EGFR-FGFR	EGFR-Src	VEGFR-Lck	PDGFR-FGFR	PDGFR-Src	Src-Lck	CDK1-CDK2	CDK1-GSK3	CDK2-GSK3	CDK1-VEGFR
	Yield of Dual-Inhibitors										
SVM	27.60	40.90	26.80	52.60	33.90	38.30	48.20	52.30	49.00	57.30	12.20
kNN	34.50	52.50	29.50	42.60	31.30	40.40	50.00	18.40	32.30	32.00	17.10
PNN	36.20	56.30	37.50	49.20	46.50	48.40	53.60	21.30	37.40	41.30	19.50
DOCK Blaster	8.60	28.20	27.70	45.90	17.00	14.40	53.60	N.A	N.A	24.50	N.A
Surflex-Dock	13.80	49.30	38.40	50.80	18.70	33.50	53.60	N.A	N.A	50.70	N.A
	False Hit Rate for Inhibitors of Other 7 Kinases Misidentified as Dual-Inhibitor of the Kinase Pair										
SVM	1.88	1.06	1.49	2.80	0.98	1.81	0.98	3.39	4.30	2.99	4.77
kNN	2.88	2.22	2.90	3.33	2.51	3.27	2.26	2.03	2.51	2.31	5.01
PNN	3.49	3.03	4.53	4.55	4.27	4.41	2.82	2.97	4.27	2.88	6.05
DOCK Blaster	25.04	33.93	35.28	34.14	25.68	19.45	37.07	na	na	40.55	na
Surflex-Dock	26.85	30.41	36.99	33.73	35.98	31.64	35.42	na	na	37.37	na
	Virtual Hit Rate for 10K Subset of 1.02 Million Zinc Clean Lead Data Set										
SVM	0.05	0.00	0.03	0.02	0.00	0.00	0.01	0.02	0.02	0.02	0.03
kNN	0.10	0.06	0.11	0.04	0.02	0.07	0.03	0.14	0.15	0.14	0.14
PNN	0.21	0.17	0.10	0.13	0.06	0.10	0.03	0.37	0.33	0.27	0.31
DOCK Blaster	0.10	0.15	0.05	0.18	0.19	0.19	0.31	na	na	0.25	na
Surflex-Dock	0.25	0.25	0.15	0.12	0.18	0.14	0.24	na	na	0.09	na
	Virtual Hit Rate for 1.02 Million Zinc Clean Lead Data Set										
SVM	0.025	0.004	0.007	0.011	0.0008	0.001	0.002	0.014	0.016	0.020	0.002
kNN	0.112	0.057	0.081	0.091	0.009	0.048	0.029	0.135	0.131	0.118	0.138
PNN	0.217	0.095	0.107	0.167	0.033	0.105	0.037	0.367	0.281	0.245	0.278
DOCK Blaster	0.141	0.247	0.158	0.236	0.291	0.144	0.348	N.A	N.A	0.389	N.A
Surflex-Dock	na	na	na	na	na	na	na	na	na	na	na

[a] Virtual hits selected by DOCK Blaster and Surflex-Dock are top 2% ranked compounds in screening 1.02 million Zinc clean lead data set or its 10K subset. The 10K subset was generated by clustering the 1.02 million Zinc clean data set into 10,000 clusters followed by the selection of one compound from each cluster as described in the text.

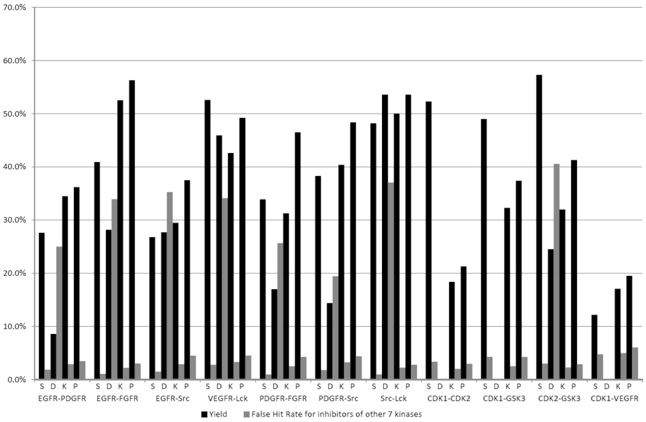

Figure 5. The comparison of the performance of C-SVM with the other three VS methods DOCK, kNN and PNN for identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3.

noted that only 12.1% of the known dual-inhibitors of the evaluated kinase pairs and 14.0% the known non-dual-inhibitors of the evaluated kinases (inhibitor of the relevant kinase but not the dual-inhibitor of the relevant kinase pair) satisfy the criteria used for assembling the Zinc-CLD. Therefore, the numbers of undiscovered dual-inhibitors in the Zinc-CLD are expected to be very small, most likely fewer than 100. Based on this estimate, the minimum and maximum numbers of false-hits produced by C-SVM, DOCK Blaster, kNN, and PNN screening of the whole-set of Zinc-CLD are 0−103 and 8−203, 1339−3863 and 1439−3963, 0−1306 and 96−1406, and 232−2730 and 332−2839 respectively. Based on the screening results of the subset of Zinc-CLD, the minimum and maximum number of false-hits of Surflex-Dock is likely to be similar or slightly better than that of DOCK Blaster. C-SVM appears to show substantially lower false-hit rates than those of the other three VS methods in screening a large compound database.

Evaluation of Combinatorial SVM Identified MDDR Virtual-Hits. C-SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. Table 6 gives the MDDR classes that contain higher percentage (≥9%) of C-SVM virtual-hits and the percentage values. We found that 58−110 or 50%−62% of the 95−175 virtual-hits belong to the antineoplastic class, which represent 0.30%−0.51% of the 21,557 MDDR compounds in the class. In particular, 34−71 or 21%−40% of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 2.9%−6.0% of the 1,181 MDDR compounds in the class. Moreover,

Table 6. MDDR Classes That Contain Higher Percentage (≥9%) of Virtual-Hits Identified by Combinatorial SVM in Screening 168 Thousand MDDR Compounds for Dual-Inhibitors of 11 Combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3

kinase pair	no. of SVM identified virtual hits	MDDR classes that contain higher % age of virtual hits	no. of virtual hits in class	percentage of class member as virtual hits
EGFR-PDGFR	175	antineoplastic	110	0.50
		tyrosine-specific protein kinase inhibitor	71	6.00
		signal transduction inhibitor	39	2.00
		antiangiogenic	25	1.50
		antiarthritic	21	0.20
EGFR-FGFR	126	antineoplastic	78	0.40
		tyrosine-specific protein kinase inhibitor	47	4.00
		antiarthritic	37	0.30
		signal transduction inhibitor	23	1.10
		antiangiogenic	16	1.00
EGFR-Src	162	antineoplastic	95	0.40
		tyrosine-specific protein kinase inhibitor	42	3.60
		signal transduction inhibitor	39	1.90
		antiangiogenic	21	1.30
		antiarthritic	15	0.10
VEGFR-Lck	170	antineoplastic	87	0.40
		antiarthritic	42	0.40
		tyrosine-specific protein kinase inhibitor	36	3.00
		signal transduction inhibitor	31	1.50
		antiangiogenic	16	1.00
PDGFR-FGFR	95	antineoplastic	58	0.30
		tyrosine-specific protein kinase inhibitor	27	2.30
		signal transduction inhibitor	22	1.10
		atherosclerosis therapy	10	0.90
		antiarthritic	10	0.10
PDGFR-Src	175	antineoplastic	103	0.50
		signal transduction inhibitor	49	2.40
		tyrosine-specific protein kinase inhibitor	40	3.40
		antiangiogenic	16	1.00
Src-Lck	131	antineoplastic	65	0.30
		tyrosine-specific protein kinase inhibitor	34	2.90
		antiarthritic	23	0.20
		signal transduction inhibitor	17	0.80
		antineoplastic enhancer	14	2.20
CDK1-CDK2	126	antineoplastic	87	0.40
		protein kinase C inhibitor	23	4.02
		antiviral	20	0.51
		tyrosine-specific protein kinase inhibitor	19	1.61
		signal transduction inhibitor	14	0.69
CDK1-GSK3	47	antineoplastic	27	0.13
		tyrosine-specific protein kinase inhibitor	10	0.85
		antihypertensive	5	0.05
		protein kinase C inhibitor	5	0.87
		antidepressant	4	0.06
CDK2-GSK3	142	antineoplastic	100	0.46
		protein kinase C inhibitor	28	4.90
		antihypertensive	21	0.19
		antiviral	20	0.51
		signal transduction inhibitor	18	0.88
CDK1-VEGFR	12	antineoplastic	5	0.02
		tyrosine-specific protein kinase inhibitor	3	0.25
		neuronal injury inhibitor	2	0.04
		antiangiogenic	2	0.12
		antiarthritic	2	0.02

13%−28% and 9%−14% of the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 0.83%−2.4% and 0.98%−1.5% of the 2,037 and 1,629 members in the two classes respectively. Therefore, many of the C-SVM 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 6.6%−29.2% false-hit rates) and inhibitors of other kinases (at 0.2%−12.7% false-hit rates).

Some of the C-SVM virtual-hits belong to the antiarthritic class. Five of our evaluated PTK-group 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.[12] VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis.[68] FGFR may partly mediates osteoarthritis.[69] PDGF-like factors stimulate the proliferative and invasive phenotype of rheumatoid arthritis synovial connective tissue cells.[70] Lck inhibition leads to immunosuppression and has been explored for the treatment of rheumatoid arthritis and asthma.[71] Therefore, some of the C-SVM virtual-hits in the antiarthritic class may be inhibitors of our evaluated kinases or their kinase-likes capable of producing antiarthritic activities.

Some of the C-SVM virtual-hits for PDGFR-FGFR belong to the atherosclerosis therapy class. Both kinases have been implicated in atherosclerosis. PDGF drives pathological mesenchymal responses in such vascular disorders as atherosclerosis, restenosis, pulmonary hypertension, retinal diseases, and fibrotic diseases.[72] Multiple FGFRs are elevated in atherosclerotic lesions in apoE−/− mice, and active FGFR-1 signaling promotes atherosclerosis development via increased SMC proliferation and by augmenting macrophage accumulation via increased expression of MCP-1 and factors promoting macrophage retention in lesions.[73] Therefore, some of the C-SVM virtual-hits in the atherosclerosis therapy may be the inhibitors of the two kinases.

Moreover, a substantial number of the C-SVM virtual-hits for the CDK1-CDK2 and CDK2-GSK3 pairs belong to the PKC inhibitor class. It has been reported that the catalytic domains of PKC and CDK share considerable homology, and inhibitors that act by blocking the catalytic site nonspecifically may act as inhibitors of both CDK and PKC,[74] which is supported by the identification of dual-inhibitors of CDK and PKC.[75,76] A number of dual-inhibitors of GSK3 and PKC have been identified.[75,77] Therefore, some of the C-SVM identified virtual-hits may be PKC inhibitors.

Does Combinatorial SVM Select Kinase Inhibitors or Membership of Compound Families? To further evaluate whether C-SVM identify kinase inhibitors rather than membership of certain compound families, compound family distribution of the identified dual-inhibitors of the 7 intra-PTK group kinase pairs was analyzed. As shown in Table 4, 15.5%, 8.5%, 11.6%, 13.1%, 15.2%, 16.0% and 16.1% of the identified EGFR-PDGFR, EGFR-FGFR, EGFR-Src, VEGFR-Lck, PDGFR-FGFR, PDGFR-Src, and Src-Lck dual-inhibitors are outside the families that contain at least one pair of non-dual-inhibitors of the two kinases of the kinase pair (i.e., at least one inhibitor for kinase A and one inhibitor for kinase B). For those families that contain at least one pair of non-dual-inhibitors of the two kinases of a kinase pair, 17.2%−68.2% of the compounds (>40.0% in majority cases) in each of these families were predicted as non-dual-inhibitors by C-SVM. These results suggest that C-SVM identify dual-inhibitors not solely based on membership to certain compound families.

Molecular Features Important for Selecting Dual-Kinase Inhibitors. The molecular features important for selecting dual-kinase inhibitors were preliminarily analyzed by testing the VS performance with varying sets of molecular descriptors. Our analysis suggested that the VS performance is critically dependent on a proper combination of multiple simple molecular property descriptors that reflect ring and hydrogen binding features, chemical property descriptors that represent hydrophobicity and molecular polarizability, molecular connectivity and shape profile descriptors that define the structural and flexibility features, and electrotopological state descriptors that determine the molecular skeletons, structural frameworks and their electronic properties. Our analysis is consistent with the reported structural analysis of the inhibitors of CDK1 and VEGFR that shows the importance of molecular structures for making extensive van der Waals contacts, hydrogen bonding with specific residues

(68) Carvalho, J. F.; Blank, M.; Shoenfeld, Y. Vascular endothelial growth factor (VEGF) in autoimmune diseases. *J. Clin. Immunol.* **2007**, *27* (3), 246–56.

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

(70) Remmers, E. F.; Sano, H.; Wilder, R. L. Platelet-derived growth factors and heparin-binding (fibroblast) growth factors in the synovial tissue pathology of rheumatoid arthritis. *Semin. Arthritis Rheum.* **1991**, *21* (3), 191–9.

(71) Meyn, M. A.; Smithgall, T. E. Small molecule inhibitors of Lck: the search for specificity within a kinase family. *Mini-Rev. Med. Chem.* **2008**, *8* (6), 628–37.

(72) Andrae, J.; Gallini, R.; Betsholtz, C. Role of platelet-derived growth factors in physiology and medicine. *Genes Dev.* **2008**, *22* (10), 1276–312.

(73) Raj, T.; Kanellakis, P.; Pomilio, G.; Jennings, G.; Bobik, A.; Agrotis, A. Inhibition of fibroblast growth factor receptor signaling attenuates atherosclerosis in apolipoprotein E-deficient mice. *Arterioscler. Thromb. Vasc. Biol.* **2006**, *26* (8), 1845–51.

(74) Kaubisch, A.; Schwartz, G. K. Cyclin-dependent kinase and protein kinase C inhibitors: a novel class of antineoplastic agents in clinical development. *Cancer J.* **2000**, *6* (4), 192–212.

(75) Shen, L.; Prouty, C.; Conway, B. R.; Westover, L.; Xu, J. Z.; Look, R. A.; Chen, X.; Beavers, M. P.; Roberts, J.; Murray, W. V.; Demarest, K. T.; Kuo, G. H. Synthesis and biological evaluation of novel macrocyclic bis-7-azaindolylmaleimides as potent and highly selective glycogen synthase kinase-3 beta (GSK-3 beta) inhibitors. *Bioorg. Med. Chem.* **2004**, *12* (5), 1239–55.

(76) Trujillo, J. I.; Kiefer, J. R.; Huang, W.; Thorarensen, A.; Xing, L.; Caspers, N. L.; Day, J. E.; Mathis, K. J.; Kretzmer, K. K.; Reitz, B. A.; Weinberg, R. A.; Stegeman, R. A.; Wrightstone, A.; Christine, L.; Compton, R.; Li, X. 2-(6-Phenyl-1H-indazol-3-yl)-1H-benzo[d]imidazoles: design and synthesis of a potent and isoform selective PKC-zeta inhibitor. *Bioorg. Med. Chem. Lett.* **2009**, *19* (3), 908–11.

(77) Zhang, H. C.; White, K. B.; Ye, H.; McComsey, D. F.; Derian, C. K.; Addo, M. F.; Andrade-Gordon, P.; Eckardt, A. J.; Conway, B. R.; Westover, L.; Xu, J. Z.; Look, R.; Demarest, K. T.; Emanuel, S.; Maryanoff, B. E. Macrocyclic bisindolylmaleimides as inhibitors of protein kinase C and glycogen synthase kinase-3. *Bioorg. Med. Chem. Lett.* **2003**, *13* (18), 3049–53.

in both kinases, and structural flexibility to accommodate the different binding site geometry and to allow the formation of alternative hydrogen bonds.[60] Our analysis is also consistent with another report that dual-kinase binding may require a combination of structural flexibility and the favorable hydrophobic interactions at specific pocket conserved in both kinase classes.[61] Moreover, many dual-inhibitors adopt specific scaffolds, such as those illustrated in Figure 3, that enable them to more easily fit to the particular regions of the ATP site,[78−80] which may be partly captured by the electrotopological state descriptors. A more comprehensive analysis using structural-based and feature selection methods[55,59] may shed more light on the detailed molecular features of dual-kinase inhibition as well as single kinase inhibition.

Concluding Remarks

Combinatorial SVM 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 combinatorial SVMs and other VS tools in identifying multikinase inhibitors and other multitarget agents may be further enhanced by incorporating knowledge of multitarget agents into VS tool development processes. With the discovery of an increasing number of selective multitarget agents from the current and future drug discovery efforts,[9,10] it is possible to introduce more comprehensive elements of distinguished structural and physicochemical features of selective multitarget agents into the training of combinatorial VS tools for more effective identification of selective multitarget agents. These multitarget VS tools may be combined with structure-based filters for enhanced target selectivity.[81] Because of their high computing speed and generalization capability, combinatorial SVM 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 multikinase inhibitors and other multitarget agents.

Abbreviations Used

VS, virtual screening; SVMs, support vector machines; C-SVM, combinatorial support vector machines; QSAR, quantitative structure−activity relationship; EGFR, epidermal growth factor receptor; VEGFR, vascular endothelial growth factor receptor; PDGFR, platelet-derived growth factor receptor; Lck, lymphocyte-specific protein tyrosine kinase; FGFR, fibroblast growth factor receptor; MDDR, MDL Drug Data Report.

Acknowledgment. This work was supported in part by grants from Singapore Academic Research Fund R-148-000-083-112, National Natural Science Foundation of China Grant 30772651, and Ministry of Science and Technology, 863 Hi-Tech Program Grant 2006AA020400.

Supporting Information Available: Summary of kNN and PNN methods and software used in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

MP100179T

(78) Petrov, K. G.; Zhang, Y. M.; Carter, M.; Cockerill, G. S.; Dickerson, S.; Gauthier, C. A.; Guo, Y.; Mook, R. A., Jr.; Rusnak, D. W.; Walker, A. L.; Wood, E. R.; Lackey, K. E. Optimization and SAR for dual ErbB-1/ErbB-2 tyrosine kinase inhibition in the 6-furanylquinazoline series. *Bioorg. Med. Chem. Lett.* **2006**, *16* (17), 4686–91.

(79) Smaill, J. B.; Baker, E. N.; Booth, R. J.; Bridges, A. J.; Dickson, J. M.; Dobrusin, E. M.; Ivanovic, I.; Kraker, A. J.; Lee, H. H.; Lunney, E. A.; Ortwine, D. F.; Palmer, B. D.; Quin, J., 3rd; Squire, C. J.; Thompson, A. M.; Denny, W. A. Synthesis and structure-activity relationships of N-6 substituted analogues of 9-hydroxy-4-phenylpyrrolo[3,4-c]carbazole-1,3(2H,6H)-diones as inhibitors of Wee1 and Chk1 checkpoint kinases. *Eur. J. Med. Chem.* **2008**, *43* (6), 1276–96.

(80) Egert-Schmidt, A. M.; Dreher, J.; Dunkel, U.; Kohfeld, S.; Preu, L.; Weber, H.; Ehlert, J. E.; Mutschler, B.; Totzke, F.; Schachtele, C.; Kubbutat, M. H.; Baumann, K.; Kunick, C. Identification of 2-anilino-9-methoxy-5,7-dihydro-6H-pyrimido[5,4-d][1]benzazepin-6-ones as dual PLK1/VEGF-R2 kinase inhibitor chemotypes by structure-based lead generation. *J. Med. Chem.* **2010**, *53* (6), 2433–42.

(81) Crespo, A.; Zhang, X.; Fernandez, A. Redesigning kinase inhibitors to enhance specificity. *J. Med. Chem.* **2008**, *51* (16), 4890–8.

(82) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.

(83) Miller, K. J. Additive Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533–42.

(84) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–8.

(85) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–45.