1-1-2012

Discovery of Novel Glycogen Synthase Kinase-3beta Inhibitors: Molecular Modeling, Virtual Screening, and Biological Evaluation

Gang Fu
University of Mississippi

Recommended Citation

Fu, Gang, "Discovery of Novel Glycogen Synthase Kinase-3beta Inhibitors: Molecular Modeling, Virtual Screening, and Biological Evaluation" (2012). *Electronic Theses and Dissertations*. 1448.
https://egrove.olemiss.edu/etd/1448

DISCOVERY OF NOVEL GLYCOGEN SYNTHASE KINASE-3β INHIBITORS: MOLECULAR MODELING, VIRTUAL SCREENING, AND BIOLOGICAL EVALUATION

A Dissertation
Presented for the Doctor of Philosophy Degree
in the Department of Medicinal Chemistry
The University of Mississippi

by

GANG FU

May 2012

ABSTRACT

Glycogen synthase kinase-3 (GSK-3) is a multifunctional serine/threonine protein kinase which is engaged in a variety of signaling pathways, regulating a wide range of cellular processes. Due to its distinct regulation mechanism and unique substrate specificity in the molecular pathogenesis of human diseases, GSK-3 is one of the most attractive therapeutic targets for the unmet treatment of pathologies, including type-II diabetes, cancers, inflammation, and neurodegenerative disease. Recent advances in drug discovery targeting GSK-3 involved extensive computational modeling techniques. Both ligand/structure-based approaches have been well explored to design ATP-competitive inhibitors. Molecular modeling plus dynamics simulations can provide insight into the protein-substrate and protein-protein interactions at substrate binding pocket and C-lobe hydrophobic groove, which will benefit the discovery of non-ATP-competitive inhibitors.

To identify structurally novel and diverse compounds that effectively inhibit GSK-3β, we performed virtual screening by implementing a mixed ligand/structure-based approach, which included pharmacophore modeling, diversity analysis, and ensemble docking. The sensitivities of different docking protocols to the induced-fit effects at the ATP-competitive binding pocket of GSK-3β have been explored. An enrichment study was employed to verify the robustness of ensemble docking compared to individual docking in terms of retrieving active compounds from a decoy dataset. A total of 24 structurally diverse compounds obtained from the virtual screening experiment underwent biological validation. The bioassay results showed that 15 out of the 24

hit compounds are indeed GSK-3β inhibitors, and among them, one compound exhibiting sub-micromolar inhibitory activity is a reasonable starting point for further optimization.

To further identify structurally novel GSK-3β inhibitors, we performed virtual screening by implementing another mixed ligand-based/structure-based approach, which included quantitative structure-activity relationship (QSAR) analysis and docking prediction. To integrate and analyze complex data sets from multiple experimental sources, we drafted and validated hierarchical QSAR, which adopts a multi-level structure to take data heterogeneity into account. A collection of 728 GSK-3 inhibitors with diverse structural scaffolds were obtained from published papers of 7 research groups based on different experimental protocols. Support vector machines and random forests were implemented with wrapper-based feature selection algorithms in order to construct predictive learning models. The best models for each single group of compounds were then selected, based on both internal and external validation, and used to build the final hierarchical QSAR model. The predictive performance of the hierarchical QSAR model can be demonstrated by an overall R^2 of 0.752 for the 141 compounds in the test set. The compounds obtained from the virtual screening experiment underwent biological validation. The bioassay results confirmed that 2 hit compounds are indeed GSK-3β inhibitors exhibiting sub-micromolar inhibitory activity, and therefore validated hierarchical QSAR as an effective approach to be used in virtual screening experiments.

We have successfully implemented a variant of supervised learning algorithm, named multiple-instance learning, in order to predict bioactive conformers of a given molecule which

are responsible for the observed biological activity. The implementation requires instance-based embedding, and joint feature selection and classification. The goal of the present project is to implement multiple-instance learning in drug activity prediction, and subsequently to identify the bioactive conformers for each molecule. The proposed approach was proven not to suffer from overfitting and to be highly competitive with classical predictive models, so it is very powerful for drug activity prediction. The approach was also validated as a useful method for pursuit of bioactive conformers.

DEDICATION

This dissertation is dedicated to everyone who guided and helped me during my pursuit of the degree, especially to my beloved family.

ACKNOWLEDGMENTS

Thanks to my research advisor and mentor, Dr. Robert J. Doerksen, Associate Professor, Department of Medicinal Chemistry, for his excellence guidance, enormous support, and consistent encouragement. His distinguished mentoring has helped me to develop well my computational, presentation, and writing skills.

Thanks to Dr. Stephen J. Cutler, Professor & Chair, Department of Medicinal Chemistry, for his continuous support and helpful suggestions. His leadership has prompted the competitive research environment and motivated me to conduct high quality research.

Thanks to Dr. Abby L. Parrill, Adjunct Professor, Department of Medicinal Chemistry, Professor and Chair, Department of Chemistry, University of Memphis, for her kindness and constructive comments. Her wide knowledge on computational chemistry has been of great value for me.

Thanks to Dr. Yixin Chen, Associate Professor, Department of Computer and Information Science, for his friendly advice and kindly guidance. His enthusiasm on science and research has remarkably inspired me.

Thanks to Dr. Christopher R. McCurdy, Dr. John M. Rimoldi, Dr. John S. Williamson, Dr. Mitchell A. Avery, and Dr. Dawn E. Wilkins, for their help and support. Their teaching with passion has helped to build up my scientific knowledge and logical way of thinking.

Thanks to Mrs. Olivia R. Dale for her help and efforts on biological evaluation of our virtual screening hit compounds.

Thanks to Mr. David Watson for his kind help on computer administration.

Thanks to funding agencies:

National Science Foundation, Mississippi EPSCoR (EPS-0903787 and EPS-0556308);

National Institute of Health’s National Center for Research Resources (COBRE: P20 RR-021929 and Research Facilities Improvements Program C06 RR-14503-01);

The center of Research Excellence in Natural Products Neuroscience (CORE-NPN);

Graduate School Council Research Grant.

Thanks to all the past and present Doerksen Research Group members, Dr. Prasanna Sivaprakasam, Dr. Pankaj R. Daga, Dr. Aihua Xie, Dr. Ronak Patel, Dr. Haining Liu, Khaled Elokely, Pankaj Pandey, Manal Nael, and Shuneize Slater, for their valuable suggestions and help.

Thanks to my friends for making my life enjoyable at University of Mississippi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. INTRODUCTION TO GLYCOGEN SYNTHASE KINASE-3 AND RELEVANT DRUG DISCOVERY

1.1. PHYSIOLOGICAL FUNCTION

Glycogen synthase kinase-3 (GSK-3) is a multifunctional serine/threonine protein kinase which is ubiquitously involved in the regulation of a wide range of cellular functions, including glucose metabolism, neuronal processes, chronic inflammations, cell proliferation and apoptosis.[1, 2] Its involvement in many cellular processes is derived from the fact that GSK-3 plays an important role in a variety of signaling pathways, most importantly Wnt signaling and insulin signaling pathway.[3] The Wnts are a family of protein ligands that influence cell proliferation, differentiation, and migration. Wnt signal transduction ultimately results in the activation of genes regulated by transcription factors, and the activation should be realized by the binding of transactivator β-catenin to the transcription factors. GSK-3 phosphorylates the N-terminal domain of β-catenin, resulting in ubiquitylation and proteasomal degradation of β-catenin. So GSK-3 plays a key inhibitory role in the Wnt pathway, which is crucial for the specification of cell fate during embryonic development.[4] The level of blood glucose is largely determined by the rate at which glucose is converted into glycogen by glycogen synthase, which is one of the important GSK-3 substrates.

1.2. REGULATION MECHANISM

It is interesting that the phosphorylation of glycogen synthase in insulin signaling pathway and β-catenin in Wnts signaling pathway are regulated through different mechanisms: insulin-induced inactivation of GSK-3 involves the phosphorylation of the serine residue in the glycine-rich N-terminal domain, whereas Wnts-induced inhibition of GSK-3 relates to the protein complex formation and displacement.[1, 5, 6] The X-ray crystal structures revealed a phosphate binding site adjacent to the active site, which constitutes three positively charged residues (Arg96, Arg180 and Lys205) to bind the priming phosphate at P+4 position of the substrate S/T-X-X-X-S/T(p) motif.[7-9] This binding stabilizes the active conformation of the activation loop, which explains the primed substrate specificity of GSK-3 and suggests a mechanism for inhibitory serine phosphorylation (so-called autoinhibition).[9] Insulin signaling promotes the phosphorylation of Ser21 in GSK-3α and Ser9 in GSK-3β near N-terminus, which transforms the N-terminus into a pseudo-substrate inhibitor that competitively occupies the same binding site and blocks the access for true primed substrate.[7] In contrast, the phosphorylation of β-catenin in Wnts signal-transduction pathway is regulated through a different mechanism. The multiprotein complex consisting of GSK-3β, axin, and adenomatous polyposis coli (APC) protein is responsible for the phosphorylation of β-catenin, and thereby promotes its ubiquitylation and destruction. The Wnts triggers a signal-transduction pathway that involves the displacement of axin-APC scaffold with FRAT (frequently rearranged in advanced T-cell lymphomas), which leads to the dephosphorylation of *β*-catenin. The axin- and FRAT-binding sites of GSK-3 near the C-terminal end introduced a new binding pocket responsible for protein-protein interaction.[10, 11]

1.3. ISOFORMS AND TAU HYPERPHOSPHORYLATION

There are two mammalian GSK-3 isoforms encoded in different genes: GSK-3α and GSK-3β.[12] Although they are highly homologous, with 84% overall identity and 98% identity in the catalytic domain, they are not functionally identical.[1] GSK-3β, also known as tau protein kinase I (TPK-I), has significant involvement in tau protein hyperphosphorylation, which has been observed in many neurodegenerative disorders such as Alzheimer’s disease (AD).[13] A new splice isoform, which contains a 13-residue insert within the kinase domain, has also been identified.[14] Analysis of the kinase activity revealed that the new splice isoform GSK-3β2 has reduced tau protein phosphorylation compared with GSK-3β1. Since GSK-3β is highly expressed in brain and is relevant to a variety of neurological disorders, it has attracted significant attention as a therapeutic target and as a molecular tool to understand the pathogenesis of these disorders.

The disease association with AD was established when GSK-3β was isolated from brain extracts and shown to produce paired helical filament (PHF) epitopes on tau. Tau is a microtubule-associated protein expressed throughout the central nervous system (CNS), but predominantly in neuronal axons. Partially phosphorylated tau contains sequence motifs that promote association with tubulin, which leads to stabilization of microtubules. However, pathological hyperphosphorylation of these motifs prevents tubulin binding and thereby results in the destabilization of microtubules.[15] There is strong evidence that GSK-3β co-localizes preferentially with insoluble neurofibrillary tangles and contributes to the formation of PHF in AD brain.[16] GSK-3β has been shown to hyperphosphorylate tau both in transfected mammalian neuronal cells and *in vivo*.[17-19]

1.4. STRUCTURAL INFORMATION

The crystal structure of GSK-3β (Figure 1.1) consists of two domains: the N-terminal domain (N-lobe) and C-terminal domain (C-lobe), and the two domains form an in-between cleft which is the ATP-binding pocket. The conserved Asp200-Phe201-Gly202 (DFG) motif at the N-lobe of the activation loop (A-loop) is in the active conformation (DFG-in conformation), where Asp200 coordinates γ-phosphate of ATP in the proper position for phosphate transfer and Phe201 makes hydrophobic contacts with the Met101 from αC-helix and the His179 in the conserved His179-Arg180-Asp181 (HRD) motif. The highly conserved Asp181 in the HRD motif at the catalytic loop is responsible for the correct configuration of the P-site serine/threonine in the peptide substrate, and most likely serves as the catalytic base to accept the proton from the hydroxyl group of the substrate serine/threonine in a proposed dissociative phosphorylation mechanism.[20] The tight electrostatic interaction between the Glu97 from αC-helix and the Lys85 from β3-strand at N-lobe generates a lobe closure which is important in the active conformation. This polar contact combined with the hydrophobic spine consisting of Leu112 at N-lobe, Met101 at αC-helix, Phe201in DFG motif, and His179 in HRD motif are responsible for the activated GSK-3 structure.[21] The triad of basic residues consisting of Arg96 from αC-helix, Arg180 from catalytic loop, and Lys205 from A-loop forms a positively charged binding pocket to accommodate the priming phosphate, which is responsible for the unique substrate specificity. The C-lobe hydrophobic groove formed by αG-helix (Gly262-Leu273) and an extended loop (Asn285-His299) presents an interface for protein-protein interaction.

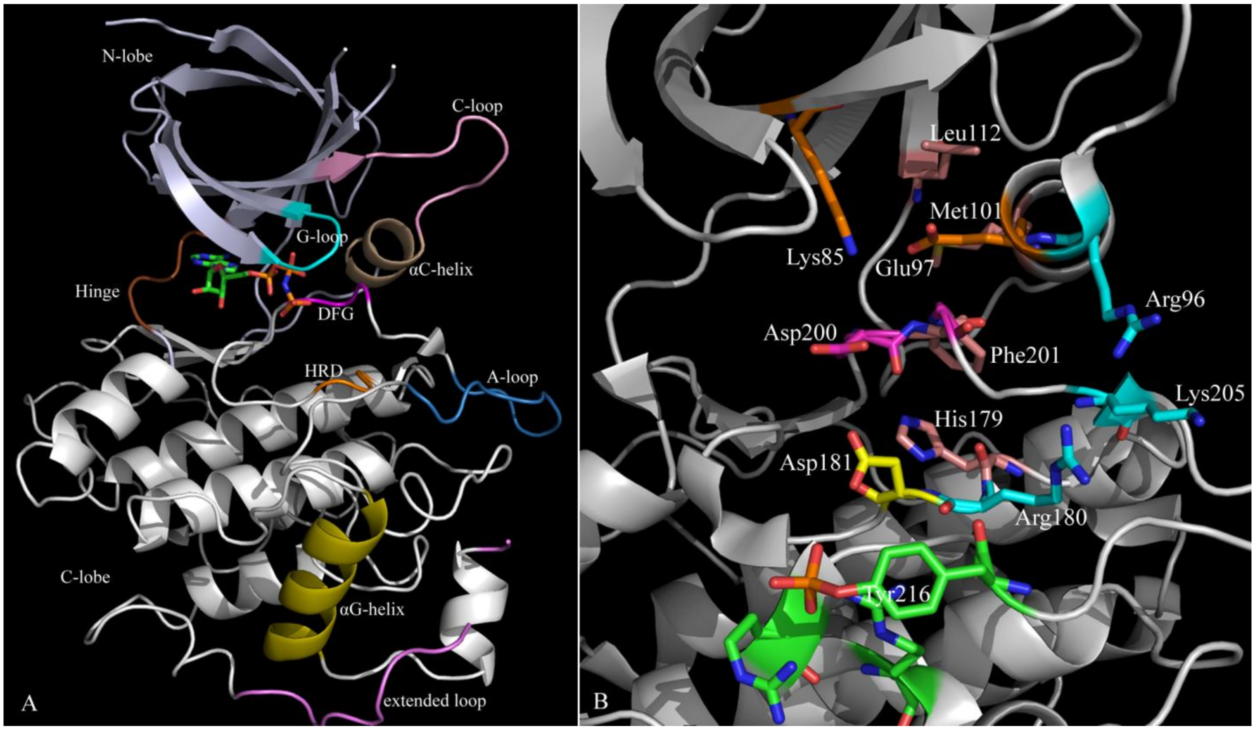


Figure 1.1. A) cartoon representation of a crystal structure (PDB code: 1PYX) of GSK-3. Color codes: light blue for N-lobe; white for C-lobe; brown for hinge region; cyan for glycine rich loop (G-loop); pink for C-loop; wheat for αC-helix; magenta for DFG moiety; orange for HRD moiety; marine for activation loop (A-loop); yellow for αG-helix; violet for extended loop; Small

molecule ADPPNP is in stick representation; B) Important residues highlighted in stick representation (PDB code: 1O9U).

1.5. DRUG DISCOVERY TARGETING GSK-3

Due to its distinct regulation mechanism and unique substrate specificity in the molecular pathogenesis of human diseases, GSK-3 is one of the most attractive therapeutic targets for the treatment of unmet pathologies, including type-II diabetes, cancers, inflammation, and neurodegenerative disease.[2, 22] The inhibition of GSK-3 phosphorylation can promote the conversion of glucose to glycogen, overcoming the resistance to insulin, which may be beneficial for the treatment of type-2 diabetes. The involvement of GSK-3 in cellular signaling pathways makes it essential in cell apoptosis and survival. The neuropathological characteristics of AD are defined by the presence of intracellular neurofibrillary tangles (NFTs) and extracellular amyloid plaques. NFTs are insoluble accumulations of hyperphosphorylated tau in the filamentous form, and amyloid plaques are dense deposits of β-amyloid (Aβ) peptides metabolized from β-amyloid precursor protein (APP). GSK-3 contributes to tau hyperphosphorylation and regulates APP processing, and inhibition of GSK-3 attenuates tau phosphorylation and Aβ levels.[23]

The aberrant regulation of GSK-3 in a variety of human diseases stimulated the development of selective and potent GSK-3 inhibitors as promising new drug candidates with great therapeutic potentials.[24] Numerous research efforts both in academy and pharmaceutical companies have shown solid evidence of preclinical and clinical efficacy for these new drug candidates in the modulation of glycogen metabolism, gene transcription and neurodegeneration.[23, 25] Most of these compounds are ATP-competitive inhibitors and none of them have demonstrated isoform selectivity. Several successful representatives (Figure 1.2) of GSK-3 inhibitors include SB 216763 and SB 415286, reported to normalize blood glucose levels in human liver cells and induce gene expression in HEK293 cells,[26] CHIR 98023 and CHIR

99021, reported to promote the activation of glycogen synthase and stimulate glycogen deposition in the liver,[27] and cazpaullones (1-azakenpaullone derivatives), reported to stimulate pancreatic *β*-cell replication and protection in isolated rat islets.[28] Those compounds might be useful for the treatment of diabetes.[25] Other useful GSK-3 inhibitors include 6-bromoindirubin-3'-oxime, which reduces the β-catenin phosphorylation on a GSK-3 specific site,[29] hymenialdisine, which blocks the phosphorylation of the microtubule-binding protein tau,[30] and AR-A014418, (aminothiazole) which was also shown to inhibit tau phosphorylation at a GSK-3 specific site.[31] Those compounds could be developed as anti-cancer agents[32] or neuroprotective agents.[33]

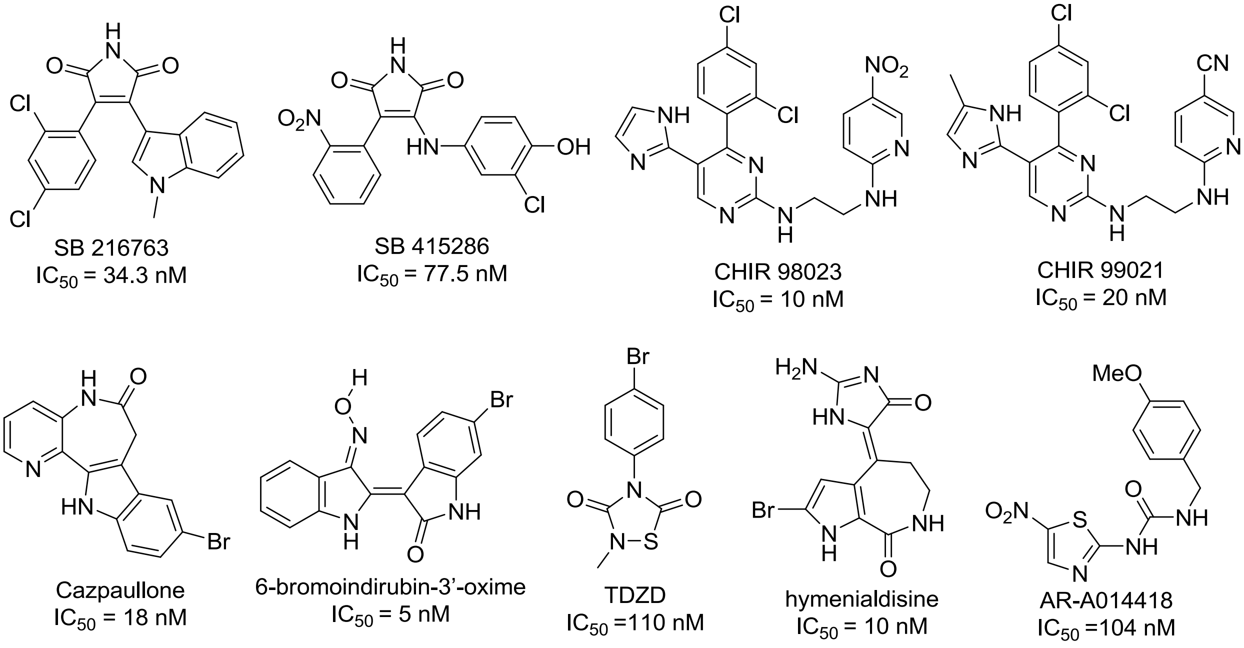


Figure 1.2. Representatives of small molecule GSK-3 inhibitors advanced into preclinical or clinical trials.

Since all kinase enzymes share a common binding site for ATP and most of the current GSK-3 inhibitors competitively interact with the ATP-binding site, design of potent inhibitors with high degrees of selectivity during drug discovery remains a challenge. To better address selectivity issues, kinase inhibitors without direct interaction with ATP-binding site provide promise of therapeutic interventions with fewer off-target side effects. A class of non-ATP-competitive GSK-3β inhibitors that have been reported as new disease-modifying agents for the effective treatment of AD and other tauopathies are thiadiazolidinones (TDZD) and derivatives (Figure 1.2).[34] Although their structure-activity relationships have been studied,[35] a clear understanding of the binding mode and inhibition mechanism is still unavailable. The substrate-competitive inhibitor L803-mts (N-Myristol-GKEAPPAPPQS(p)P) has demonstrated promising preclinical merit through *in vivo* inhibition of GSK-3, including antidepressant-like activity based on the evidence of up-regulated β-catenin in mouse hippocampus;[36] insulin mimetic action based on the facts of elevated glycogen synthase activity and increased glucose uptake.[37] Furthermore, long-term administration of L803-mts into mice can reduce blood glucose levels, improve glucose tolerance and homeostatis in a diabetic model.[38]

Although a number of GSK-3 inhibitors have emerged and several of them are fairly potent, none of them have been developed as effective drug candidates nor approved by the US Food and Drug Administration. The major reasons frustrating all the efforts are kinase selectivity issues and poor pharmacokinetic profile including factors such as CNS bioavailability. Hence there is still a great need to identify and develop structurally novel and diverse GSK-3β inhibitors as potential therapeutic interventions. Various efforts have been made in the discovery and development of potent and selective GSK-3 inhibitors and the most fruitful one has been computer-aided drug design (CADD), which accelerated the lead evolution and optimization for the pursuit of structurally novel and diverse GSK-3 inhibitors. CADD approaches can typically be divided into two classes: ligand-based approaches and structure-based approaches. For ligand-based approaches, the inhibitory activities of thousands of compounds against GSK-3 have been reported, along with extensive quantitative structure-activity relationship (QSAR) studies. QSAR has been employed to correlate the biological activities with the structural or physicochemical properties, and the correlations subsequently provided the distinguishing and favorable characteristics. For structure-based approaches, around 30 crystal structures of GSK-3 currently have been deposited in Protein Data Bank (PDB[39]) (1GNG,[11] 1H8F,[7] 1I09,[8] 1J1B,[40] 1J1C,[40] 1O9U,[10] 1PYX,[41] 1Q3D,[41] 1Q3W,[41] 1Q41,[41] 1Q4L,[41] 1Q5K,[42] 1R0E,[43] 1UV5,[44] 2O5K,[45] 2OW3,[46] 2JLD,[47] 3DU8,[48] 3F7Z,[49] 3F88,[49] 3GB2,[50] 3I4B,[51] 3L1S,[52] 3M1S,[53] and 3PUP[54]), which were utilized to predict the binding modes and affinities of new inhibitors.

Several successful high throughput virtual screening experiments were conducted via ligand-based approaches, especially quantitative structure-activity relationship (QSAR) analyses and pharmacophore screening.[55-57] The reason that previous researchers preferred ligand-based virtual screening rather than structure-based virtual screening is that the X-ray crystal structures only provide a static picture of ligand-protein complexes and give limited information as to how protein flexibility can be exploited for the purpose of drug discovery. Docking and enrichment studies in the adaptive ATP-binding site of six GSK-3β crystal structures have shown the poor prediction accuracies of docking poses without considering the significant induced fit effects.[58] However, structure-based approaches not only give us information regarding the best possible fit of a molecule in the binding site, but also provide insight into the important binding features essential to the ligand-protein interaction that can be used to address selectivity problems especially against highly homologous kinases.[59, 60]

Chapter 2. RECENT ADVANCES IN COMPUTATIONAL MODELING TARGETING GSK-3 FOR DRUG DISCOVERY

2.1. COMPUTATIONAL MODELING TARGETING ATP-BINDING POCKET

2.1.1. LIGAND-BASED APPROACHES

2.1.1.1. QSAR ANALYSIS

Extensive QSAR studies have been applied on GSK-3 inhibitors (Appendix: A), including 3-anilino-4-arylmaleimides,[61] indirubins,[44, 62, 63] paullones, [64] aloisines,[65] 2,4-disubstituted thiadiazolidinones (TDZD),[34, 35] pyrazolopyrimidines,[66, 67] pyrazolopyridazine,[68, 69] pyrazolopyridines,[70-72] and benzofuran-3-yl-(indol-3-yl)maleimides.[73] Exploration of QSAR entails statistically significant correlation between structural and physicochemical properties (so-called independent variables) and biological activities (so-called dependent variables) of the chemical structures.

Katritzky *et al.* reported a 2D-QSAR study of 277 GSK-3 inhibitors using geometrical, topological, quantum mechanical, and electronic descriptors.[74] The study compared both a linear QSAR method using multiple linear analysis (MLA) and a nonlinear QSAR method using artificial neural networks (ANN). Based on internal validation, the MLA produced highly predictive models for 3-anilino-4-arylmaleimides, moderately predictive models for pyrazolopyridines, and weakly predictive models for pyrazolopyridazines and pyrazolopyrimidines. In comparison, nonlinear QSAR modeling of the whole collection of compounds using ANN yielded acceptable predictions for both training set and test set. However, the interpretability was compromised in the nonlinear model. Sivaprakasam *et al.* also reported a 2D-QSAR study using Fujita-Ban and Hansch analysis, which explored the physicochemical and structural requirements for a set of 3-anilino-4-phenylmaleimides toward

GSK-3α binding.[75] The interpretability of Fujita-Ban and Hansch analysis is high and they agreed on the conclusion that hydrophobic interaction at the 3-anilino ring as well as steric and electronic interactions on the 4-phenyl ring are crucial for inhibitory activities.

To improve interpretability, 3D-QSAR using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) is very helpful. One of the most crucial steps for CoMFA and CoMSIA methods is structural alignment, which requires a reference as putative bioactive conformation. Sivaprakasam *et al.* reported a 3D-QSAR study using CoMFA and CoMSIA to further examine the structural requirements toward GSK-3α binding of 3-anilino-4-arylmaleimides.[76] Two structural alignment strategies were employed and compared, which were ligand-based alignment using the lowest energy conformation of the most active compound and structure-based alignment using the locally minimized conformation of co-crystallized compound. Based on statistical results, the structure-based alignment produced the best models for both CoMFA and CoMSIA. Two alignment strategies have also been employed by Zeng *et al.* to carry out CoMFA and CoMSIA for aloinsines as GSK-3 inhibitors.[77] In comparison, structure-based alignment which utilized a high energy conformation extracted from a co-crystallized structure of aloisine B with CDK2 yielded much better statistical parameters for both CoMFA and CoMSIA. So the bioactive conformation is not usually the lowest in energy. Another 3D-QSAR study using CoMFA and CoMSIA was carried out by Zhang *et al.* for indirubins, and it also compared two alignment rules.[78] However, the comparison was performed in a different way. The same template conformation of indirubin-3'-oxime extracted from a co-crystallized structure was employed, and two different sets of conformations for the compounds in the training set were used in two alignment rules for comparison. In the so-called receptor-based method, the docking poses of compounds in the training set were used for superimposition. In the so-called ligand-based method, the minimum energy conformations for the compounds in the training set were superimposed on the same template conformation of indirubin-3’-oxime. The comparison showed evidence that the minimum energy conformations produced better predictions for both CoMFA and CoMSIA. Since the differences of statistical results for the two alignment rules were not significant, both methods are suitable to build reliable 3D-QSAR models.

When the binding modes are not available from crystal structures, the 3D-QSAR analysis can be imperative to explore the structural requirements for ligand-protein binding. TDZDs are identified as the first ATP-noncompetitive GSK-3 inhibitors, which block phosphorylation without targeting ATP binding. These compounds are of great interest since they did not show inhibitory activity against other kinases and the mechanism of their inhibitory action is still not clear. A CoMFA study based on the alignment of minimum energy conformations produced a predictive model for TDZDs, which was externally validated using an independent test set.[35] 3D-QSAR can be also very important in selectivity studies to explore the correlations between the chemical structures and the multiple biological activities. Paullones exhibited multiple inhibitory activities against CDK1, CDK5, and GSK-3. Three CoMSIA models were established and compared. Since the structural alignment was based on the minimized docking poses at the CDK1 ATP-binding site, the statistical results obtained for the CDK1 model were clearly superior to the ones for the CDK5 model and the GSK-3 model. The 3D contour maps for the inhibitory activities of paullones with respect to CDK1, CDK5, and GSK-3 indicated that the electronic fields between the models should be taken into account for the development of GSK-3 selective paullones[64].

A 3D-QSAR study using CoMFA and CoMSIA followed by molecular docking can be used as a conventional approach to unveil the structural requirements for ligand-protein interactions, which is beneficial for medicinal chemists to optimize lead compounds with high interpretability. However, the structural alignment of 3D-QSAR requires the molecules to have a similar scaffold, which typically limits the predictiveness of the models especially for a large dataset consisting of structurally diverse compounds. Also, sometimes the conformations used for alignments are significantly different from the bioactive conformations, which can reduce the accuracy and relevance of the model. In contrast to 3D-QSAR methods, classical QSAR methods based on structural and physicochemical descriptors are independent of structural alignment, so they can be expected to perform well with large, diverse data sets. Furthermore, with the development of machine learning algorithms and artificial intelligence methods which can be implemented for both model construction and feature selection, modern QSAR methods which can produce highly predictive linear or nonlinear models play an increasingly important role in the drug discovery process, especially in virtual screening studies to identify novel hit compounds.[79]

Taha *et al.* extensively surveyed the literature and compiled a large group of diverse GSK-3 inhibitors, including 3-anilino-4-arylmaleimides,[61] pyrazolopyridazine,[68, 69] and pyrazolopyridines.[70-72] The biological activities of these compounds were obtained by testing against the human GSK-3α isoform under the same experimental protocol, so their bioactivities are comparable. Two subsets of compounds carefully selected from the collection of 152 diverse GSK-3 inhibitors were employed to explore the pharmacophoric space using the HYPOGEN module from CATALYST software package, which yielded in total 60 different pharmacophore models having satisfactory statistical results. Although pharmacophore models can be used to explain ligand-protein interactions, their predictive abilities to correlate the chemical structures to the bioactivities are limited by steric shielding and auxiliary substituent groups that can either enhance or reduce the bioactivity. So a self-consistent QSAR analysis using a genetic function algorithm and multiple linear regression (GFA-MLR-QSAR) was performed to search for the best combination of pharmacophore models and physicochemical descriptors. To reduce redundancy, the 10 best scoring pharmacophore models selected from 10 clusters of 60 models were combined with quantitative physicochemical descriptors to construct the descriptor space. GFA-MLR produced a highly predictive QSAR model, which contained two orthogonal pharmacophore models and seven different physicochemical descriptors. The optimal QSAR model obtained from GFA-MLR was subsequently employed in virtual screening to select potential GSK-3β inhibitors from an in-house-built structural database of established drugs. The top three ranked drugs, namely, hydroxychloroquine, cimetidine, and gemifloxacin, were validated to have GSK-3β inhibitory activities in both *in vitro* and *in vivo* models.[80] Goodarzi *et al.* performed a series of QSAR studies using the same set of 152 diverse GSK-3 inhibitors with the same division of training and test set.[81] Their linear and nonlinear QSAR models exhibited the powerful predictivenessof modern QSAR analysis using artificial intelligence and machine learning algorithms. The artificial intelligence algorithm named fuzzy rough set ant colony optimization combined with multiple linear regression and support vector machines yielded the best linear predictive model and the best nonlinear model, which had highpredictiveness.

Most recently, 3D-QSAR analysis has emerged as a useful post-filtering predictor that can be used to predict the bioactivities of structurally similar hit compounds obtained from docking screening. The incorporation of 3D-QSAR prediction into the virtual screening protocol has been validated to be reliable and beneficial to search for lead compounds inhibiting epidermal growth factor receptor.[82] Fang *et al.* proposed and validated a virtual screening protocol that combined structure-based and ligand-based approaches to search for new lead compounds inhibiting GSK-3β.[83] The ligand-based 3D-QSAR analysis using CoMFA and CoMSIA was performed on a data set of benzofuran-3-yl-(indol-3-yl)maleimides. The best CoMFA and CoMSIA models were then externally validated using two different test sets, and the best CoMSIA model was selected as the most predictive for the structurally diverse compounds. Then, the best CoMSIA model was combined with molecular docking in their virtual screening protocol, which yielded a hit rate greater than 20%. Finally, an enrichment study was performed to validate that the proposed virtual screening protocol using combined molecular docking and 3D-QSAR prediction was reliably able to retrieve active compounds from a virtual library. It was proven that the proposed virtual screening protocol indeed improved the hit rate by approximately 1.5 times during screening of one fifth of the compounds of the virtual library, compared with a virtual screening protocol without ligand-based 3D-QSAR prediction.

2.1.1.2. PHARMACOPHORE MODELING

Since a variety of GSK-3 inhibitors have been reported recently, information regarding the chemical structures of known GSK-3 inhibitors could be well utilized to identify novel scaffolds using a pharmacophore mapping strategy. A 3D common feature pharmacophore model unveils crucial information regarding the 3D arrangement of essential common features to be recognized by the active site during ligand-protein binding. Dessalew *et al.* carried out a pharmacophore mapping study using a set of 21 potent and structurally diverse GSK-3 inhibitors.[56] The top-ranked pharmacophore model was subsequently used as a query to screen a chemical database. Pharmacophore-based virtual screening followed by molecular docking yielded five hits with novel scaffolds, which, however, were not biologically validated.

Exploration of pharmacophoric space carried out by Taha *et al.* produced various putative pharmacophore models, which can be complementary to each other indicating distinct binding modes accessible for GSK-3 inhibitors.[80] So the validation and selection of pharmacophore models will be essential in pharmacophore-based virtual screening. Patel *et al.* developed a specific pharmacophore model for selective GSK-3 inhibitors using the distance comparison method (DISCO), which was validated by two strategies: 1) overlap of pharmacophore features on important interactions for ligand-protein binding; 2) searching a database containing selective and non-selective GSK-3 inhibitors as well as inactive molecules.[84] The two important interactions that a pharmacophore model should demonstrate included a hydrogen bond acceptor interacting with Val 135 and a hydrogen bond donor interacting with Asp 133. The validation database contained a set of 378 compounds, including 130 selective inhibitors, 216 non-selective inhibitors, and 32 inactive compounds. The specific pharmacophore model finally selected as a query for virtual screening demonstrated satisfactory discriminatory ability by picking 96 out of 130 selective inhibitors, only 5 out of 216 non-selective inhibitors, and only 2 out of 32 inactive molecules. The final specific pharmacophore query containing 8 features was selected to be used for virtual screening. The hits were docked into GSK-3β ATP-binding pocket (1Q4L), and 9 potential lead compounds were identified exhibiting high docking scores and putative docking poses. Biological validation was not performed.

Pharmacophore mapping alone is important but not sufficient to adequately account for distinct binding modes and provide effective discrimination between active and inactive compounds. So it usually is used in combination with other ligand-based approaches. For instance, Taha *et al.* proposed a self-consistent QSAR analysis searching for the best combination of pharmacophore models and physicochemical descriptors to discriminate active and inactive molecules.[80] Kim *et al.* carried out sequential ligand-based virtual screening by combining common feature pharmacophore mapping and recursive partitioning (RP) classification to identify novel GSK-3β inhibitors.[55] The pharmacophore models were derived from six known GSK-3β inhibitors and validated through evaluating hit rates in an artificial virtual screening experiment against a collection of 287 known GSK-3β inhibitors and 994 inactive compounds. The best common feature pharmacophore model provided effective discrimination between active inhibitors (hit rate=45%) and inactive compounds (hit rate=18%). An optimal RP classification model constructed using nine *E*-state keys and one topological index was applied sequentially after pharmacophore-based virtual screening to further filter the database. The final 56 hits were carefully selected considering docking pose, structural diversity, and synthetic accessibility. They were then subjected to biological validation, and three compounds exhibited low micromolar GSK-3β inhibitory activities.

2.1.2. STRUCTURE-BASED APPROACHES

2.1.2.1. VIRTUAL SCREENING AND DE NOVO DESIGN

High throughput virtual screening using molecular docking has been widely applied in structure-based drug discovery.[85] Kang *et al.* successfully identified TDZDs as submicromolar ATP-competitive GSK-3β inhibitors using structure-based virtual screening.[86] Out of 170 TDZDs, five compounds were selected based on Hammerhead docking scores and structural diversity. Most interestingly, out of five compounds that were bioassayed, the two most active compounds demonstrated ATP-competitive inhibition and high selectivity over other homologous kinases.

The direct comparison of the performance between virtual screening using molecular docking and experimental high throughput screening was first carried out by Polgár *et al.* as part of a lead discovery project.[87] Due to the conformational flexibility of Gln185 at the ATP-binding pocket (Figure 2.1) of GSK-3β crystal structures, three representative structures with different Gln185 conformations (1Q4L, 1UV5, and 1Q3D) were selected for docking studies. In accordance with the artificial enrichment study, FlexX-Pharm which incorporates pharmacophore constraints into molecular docking produced the highest enrichment factor and demonstrated an improved ability to filter out false positives in the decoy dataset. Hence, FlexX-Pharm was subsequently employed in the direct comparison of virtual and experimental high throughput screening, which was performed in a real enrichment experiment using a corporate collection of 16,299 diverse molecules. The experimental screening yielded 90 validated hit compounds (hit rate = 0.55%). The best structure-based virtual screening achieved a 23-fold improvement of enrichment factor for the top 1% of the ranked database. However, the 69 validated hit compounds could not be identified by the virtual screeing algorithm (false negatives). Even if the whole ranked database was considered, FlexX-Pharm could not generate reasonable docking poses for 49 of the validated hit compounds, so these false negative compounds would be lost by virtual screening. It was also demonstrated that the correlation between the docking scores and the inhibitory activities was extremely low, and even the distributions of the docking scores and the experimental measures were different. So there existed significant uncertainty in the prediction of inhibitory activities using structure-based virtual screening. The results suggested that virtual screening can be useful for filtering out false positives and can be complementary to the experimental screening in lead identification from a very large database.

Structure-based Ludi de novo design, which is a fragmental approach applying the principle of complementarity, has been demonstrated as a useful tool to identify structurally novel compounds predicted to be GSK-3 inhibitors.[88] The method first calculates interaction sites within the binding pocket, providing information about steric, hydrophobic, electrostatic, van der Waals, and hydrogen bonding interactions; and then searches for fragments that complement the binding sites, which are subsequently ranked based on the Ludi empirical scoring function. Dessalew *et al.* carried out de novo design experiments and identified 10 potential leads sharing a 2,4-diaminopyrimidine scaffold and 5 potential leads sharing a 2,4-diaminoquinazoline scaffold. Visual examination of the docking poses and comparative analysis of docking scores provided further confidence on the prediction, but the hits were not biologically validated.

2.1.2.2. DOCKING PREDICTION AND INDUCED FIT EFFECTS

To better understand the protein flexibility and induced fit effects in structure-based drug discovery, Gadakar *et al.* investigated the prediction accuracies of docking poses for GSK-3β inhibitors in the binding site of 6 crystal structures (1H8F, 1PYX, 1O9U, 1Q4L, 1Q5K, and 1UV5), using the Glide module from the Schrödinger suite.[58] Glide SP, Glide XP, and induced fit docking (IFD) were utilized to carry out self-docking, cross-docking, and enrichment studies. Both Glide SP and XP exhibited acceptable abilities to reproduce the co-crystallized binding poses in a self-docking experiment and Glide XP performed slightly better than Glide SP. However, most of the cross-docking predictions could not reproduce the co-crystallized binding poses, and the large deviations of the docking poses indicated significant induced fit effects in the binding of GSK-3β to various ligands. To support this observation, IFD was carried out using 3 crystal structures (1Q5K, 1O9U, and 1Q4L) to compare the cross-docking predictions. IFD, which can simulate the conformational flexibility of the protein binding site during ligand-protein recognition, consistently improved the docking predictions based on the markedly reduced RMSD values. The subsequent enrichment study also demonstrated the utility of induced fit models in the binding prediction based on the improved retrieval of active inhibitors seeded in a decoy database.

The induced fit effects are attributed to the conformational flexibility of the protein as well as the bridging water molecules at the binding site. The sensitivity of docking prediction to the presence of bridging water molecules has been well investigated, which indicated the significance of including water molecules in a docking simulation.[89] Furthermore, the displacement of water molecules with a small molecule at the binding site is a major contribution to molecular recognition, and this favorable contribution to the binding free energy has been quantitatively described by capturing the hydration map (so-called water map) of the thermodynamic properties (especially enthalpy and entropy) of the active site solvent.[90] The enthalpic and entropic contributions characterized by the expulsion of hydrophobically enclosed solvent by the complementary small molecules has been well correlated to the binding free energy differences as part of structure-activity relationship analysis.[91] Most recently, the water map of the location and energetics of the active site solvent was found to be able to explain quantitatively the kinase selectivity SAR for four pairs of kinase systems (Src/GSK-3β, Abl/c-Kit, ZAP-70/Syk, CDK2/CDK4).[92] Some of the enclosed active site water molecules constitute a conserved structural element contributing to the extended network of hydrogen bonds at the kinase ATP-binding site.[93] The examination of 13 protein kinases with active conformations has revealed the presence of conserved water molecules as essential structural elements interconnecting the protein structures and stabilizing catalytic residues.[94]

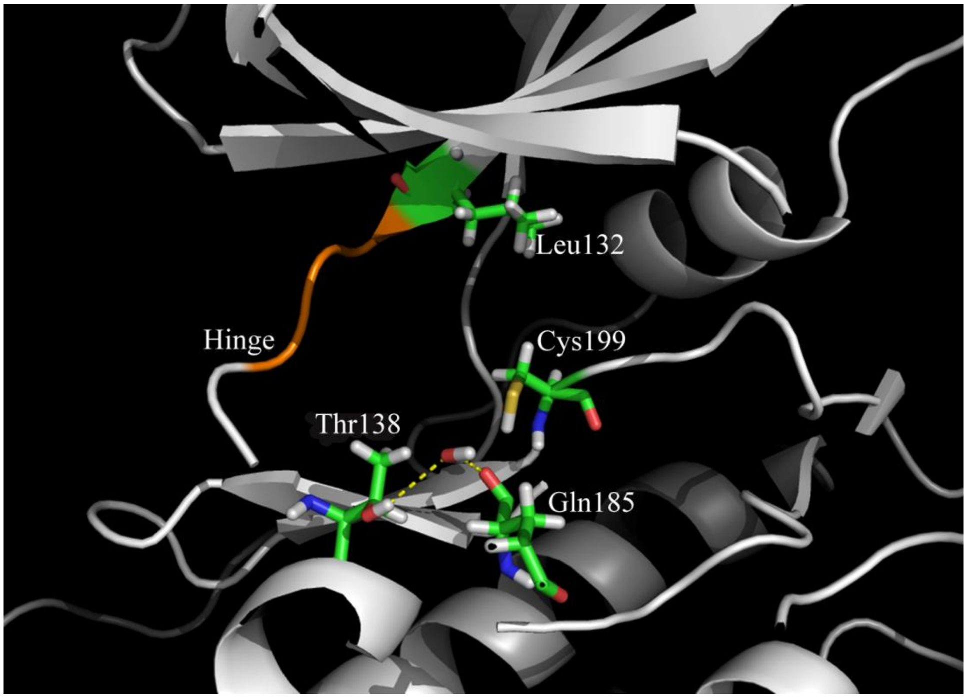


Figure 2.1. The ATP-binding pocket with important residues and conserved water molecule highlighted in stick representation; protein is in cartoon representation; hinge region is highlighted in orange (PDB code: 1Q5K).

Lu *et al.* extensively investigated the roles of conserved bridging water molecules in the binding of GSK-3β to inhibitors using 10 crystal structures of ligand-protein complexes (1Q3D, 1Q3W, 1Q41, 1Q4L, 1Q5K, 1R0E, 3DU8, 3F7Z, 3GB2, and 3I4B) [95]. ONIOM-based quantum mechanics/molecular mechanics (QM/MM) calculations were used to optimize the co-crystallized structures and identified the conserved bridging water molecules at the GSK-3β ATP-binding site that form hydrogen bonds with the side chain hydroxyl group of Thr138 and the backbone carbonyl group of Gln185 (Figure 2.1), except 1R0E for which only hydrogen bonds with Thr138 were observed. Thy fully optimized geometries of the QM layer which included inhibitors, bridging water molecules, and two residues (Thr138 and Gln185) did not undergo significant structural changes compared with the original crystal structures. The theory of atoms in molecules (AIM) was employed to examine the properties at the hydrogen bond critical points which confirmed the existence of water-mediated hydrogen bonding networks.

The subsequent molecular dynamics (MD) simulations (6 ns) were performed to compare two complex systems based on crystal structure 1R0E with and without the bridging water molecule, which demonstrated that the bridging water molecule was locked in the binding site and stabilized the protein structure via water-mediated hydrogen bonding interactions. Finally, molecular docking studies using 8 crystal structures (1Q3D, 1Q3W, 1Q41, Q4L, 1Q5K, 1R0E, 3F7Z, and 3I4B) demonstrated that the inclusion of bridging water molecules can improve both docking pose prediction and binding affinity prediction.

2.1.2.3. SELECTIVITY STUDIES

Since all the protein kinases share a common ATP-binding pocket, development of specific and selective ATP-competitive kinase inhibitors is highly challenging. Structure-based approaches may shed light on rational drug design of selective and specific kinase inhibitors using spatial information related to subtle differences at the ATP-binding pocket. Vulpetti *et al.* carried out a comparative study using GRID/CPCA and GRIND/CPCA (CPCA=consensus principal component analysis; GRIND=Grid-Independent Descriptors) on a set of 10 crystal structures including 4 complexes of CDK2/cyclin A bound to inhibitors, and 6 complexes of GSK-3β bound to inhibitors.[96] GRID/CPCA requires structural alignment and GRIND/CPCA is alignment-independent. The direct comparison of 3D structures of the ATP-binding pockets highlighted the regions and interactions useful to gain selectivity against specific targets of interest. Inclusion of multiple crystal structures took into account protein flexibility. In order to identify the most discriminative interactions between the ATP-binding pockets of CDK2 and GSK-3β, CPCA was employed to analyze the multivariate descriptions obtained from molecular interaction fields (MIFs) calculations. 3D visualization of the GRID/CPCA contour plots superimposed on the crystal structures of CDK2 with bound benzodipyrazole defined a precise spatial position of a hydrophobic site in the back of ATP-binding pocket that can be exploited to improve selectivity. The sequence and structural alignment identified the two residue differences which contribute to the discriminative regions: Phe80 in CDK2 compared to Leu132 in GSK-3β, and Ala144 in CDK2 compared to Cys199 in GSK-3β (Figure 2.1). The computational insights helped the design of a 4,4-gem-dimethyl derivative as a selective inhibitor for CDK2, which was subsequently biologically validated. GRID/CPCA was also carried out to explain the fact that a 6-bromo substituent on indirubin-3'-oxime increases the selectivity for GSK-3β over CDK2. It is interesting that the selectivity region for GSK-3β is very close to the selectivity region for CDK2. The selectivity profile can be explained by the subtle difference of hydrophobic interactions in the back of the ATP-binding pocket, while the increased width of this pocket better suited the bromine substituent to achieve selectivity for GSK-3β and the increased depth of this pocket complemented well the dimethyl groups to achieve selectivity for CDK2. GRIND/CPCA without superimposition confirmed the same selectivity regions. The good agreement of the two different analyses supports the reliability of the results. Another study has successively exploited the structural difference between Leu132 in GSK-3β and Phe80 in CDKs to design a series of 7-substituted aminoindazoles as potent GSK-3β inhibitors with high selectivity against CDK1 and CDK2.[97]

Besides the aforementioned statistical approach, MD simulations combined with binding free energy calculations and decomposition analysis can also provide insight into the selectivity profile. Molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) binding free energy calculations and molecular mechanics/generalized Born surface area (MM/GBSA) free energy decomposition analysis were carried out by Chen *et al.* to explore the selectivity profile of paullones.[98] Six ligand-protein complexes of three paullones (alsterpaullone, 1-azakenpaullone, and 2-azakenpaullone) binding to protein crystal structures (1O9U for GSK-3 and 1UNL for CDK5) were constructed. A detailed analysis of energy components contributing to the binding affinities revealed that van der Waals interactions contributed to the major favorable binding free energies. But the small variance of van der Waals contributions among the six complexes cannot explain the selectivity profiles of paullones. The sequence and structural alignment demonstrated that two parallel residues at the conserved position may distinguish the selectivity of paullones in favor of GSK-3: Val135 in GSK-3 and Cys83 in CDK5, and Tyr138 in GSK-3 and Asp86 in CDK5. A residue-based MM/GBSA decomposition analysis was carried out to calculate the interaction of each residue-ligand pair, which indicated that the net electrostatic contribution of alsterpaullone with Val135 in GSK-3 was indeed much stronger than the one of alsterpaullone with Cys83 in CDK5, and the same variances were observed for azapaullones. However, different electrostatic contributions were not useful for structural optimization to gain selectivity, since the hydrogen bond interactions occurred between the ligands and the backbone of Val135 and of Cys83. Another observation from the energy decomposition analysis can be used to explain the selectivity gain of 1-azapaullone that occurs when substituting nitrogen for carbon at the 1-position. The substitution resulted in the reduction in pairwise interaction between Asp86 and 1-azapaullone compared to that between Asp86 and alsterpaullone, which was consistent with the reduction in the occupancies of hydrogen bonds of N12 in 1-azapaullone with Asp86 compared to N12 in alsterpaullone with Asp86. In contrast to the interaction between ligands and CDK5, the substitution did not change much the interaction between ligands and Tyr138 in GSK-3. So the substitution did not change much the inhibitory activity for 1-azapaullone against GSK-3, but reduced the inhibitory activity significantly for 1-azapaullone against CDK5. Hence, the interaction toward Asp86 in CDK5 was able to be employed to improve the selectivity profile in structural optimization, especially for paullones.

2.2. COMPUTATIONAL MODELING TARGETING THE SUBSTRATE BINDING POCKET

Considering the unique substrate specificity of GSK-3, it is worthwhile to explore carefully the substrate binding pocket to design small peptide inhibitors such as L803-mts (N-Myristol-GKEAPPAPPQS(p)P), which exhibited promising preclinical values. To identify the interaction sites located within the substrate binding cleft, Ilouz *et al.* carried out molecular modeling combined with biological studies.[99] The sequence alignment identified three residues of interest at the substrate binding site: Gln89 and Asn95 in the N-lobe are conserved in GSK-3 but not conserved in other homologous kinases including mitogen-activated protein kinase (MAPK), cyclin-dependent kinase 2 (CDK2), and protein kinase A (PKA); Phe67 from the glycine-rich loop (G-loop, also known as the P-loop) is conserved in GSK-3 and PKA but is mutated to tyrosine in MAPK and CDK2. The biological studies demonstrated that the mutations of the three residues will result in various reductions in levels of GSK-3 phosphorylation. To understand the substrate recognition mechanism at the atomic level, the ternary complex structures of phosphorylated GSK-3β (pTyr216), ATP, and the substrates, the phosphorylated cAMP responsive element binding proteins (pCREBs), were constructed by protein-protein docking. The docking studies verified that the polar residues, Gln89 and Asn95, participated in hydrogen bond interactions with various polar/charged residues at the P+6 position in the substrates; and the Phe67 in the conserved G-loop pointed toward the substrate binding cleft to stabilize the G-loop conformation through hydrophobic contact with substrates. However, the two polar residues (Gln89 and Asn95) are far away from the catalytic site at the P position, so they were rarely explored for the design of potent substrate-competitive inhibitors. Figure 2.2 shows the key residues involved in the unique substrate specificity of GSK-3.

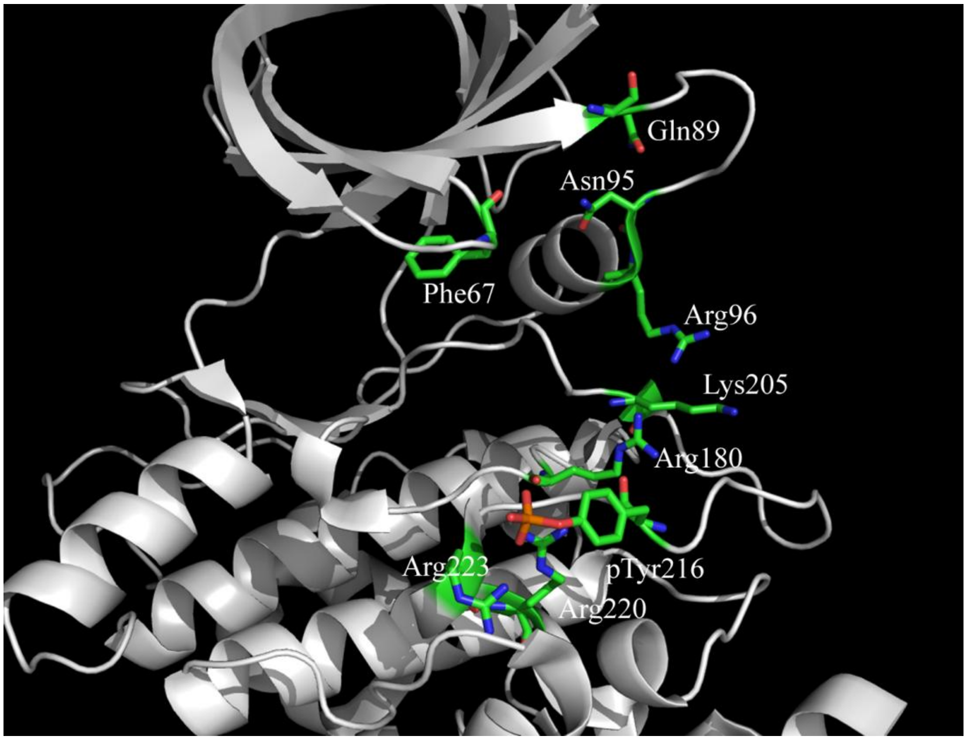


Figure 2.2. The primed substrate binding pocket with important residues highlighted in stick representation (PDB code: 1O9U).

The unique substrate specificity of GSK-3 is derived from the fact that the priming phosphate is well accommodated by three positively charged residues. The mutation of Arg96 to lysine or alanine severely impaired the phosphorylation of primed substrates without affecting non-primed substrates. To explain the mutagenesis study, Zhang *et al.* performed MD simulations on three systems: wild type (WT), R96K and R96A mutants of GSK-3β-ATP-pSer complexes, followed by MM-GBSA binding free energy analysis.[100] MD simulations demonstrated that Arg96 was important to induce a slight closure of the N- and C-lobes. The lobe closure involving a closed conformation of the C-loop, A-loop, and G-loop may facilitate substrate binding and ATP positioning. The mutation of Arg96 which is located on the αC-helix caused the open motion of the disordered C-loop and G-loop, and subsequently twisted the conformation of ATP's flexible triphosphate moiety. Abnormal conformational changes which occurred on the G-loop were related to the high mobility of β and γ phosphate groups in two mutants during MD simulations. Binding free energy analysis provided evidence that the mutation indeed reduced the binding affinities, and especially disturbed the electrostatic interaction between Arg96 and pSer, which was a dominant favorable contribution to the binding energy.

To obtain an atomic level description of the activation mechanism by Tyr216 phosphorylation, MD simulations were carried out by Buch *et al.* on both the inactive form of the unphosphorylated GSK-3β-ATP complex (derived from 1PYX) and the active form of the phosphorylated GSK-3β(p)-ATP complex (derived from 1O9U).[101] The unphosphorylated Tyr216 in the crystal structure 1O9U points into the priming phosphate binding pocket and blocks the access of the primed substrate, while the phosphorylated Tyr216 in the crystal structure 1PYX points in the opposite direction and forms polar contacts with Arg220 and Arg223 (Figure 2.2). The intramolecular electrostatic interactions at the active site were further monitored throughout MD simulations on both active and inactive forms. Due to the different orientation of Tyr216, Arg220 can form different polar contacts controlling the accessibility of the catalytic groove. In the inactive form, Arg220 interacts with the γ-phosphate group of ATP and Asp181 from the conserved HRD motif. In addition to another electrostatic interaction between Arg96 from αC-helix and Asp200 from the conserved DFG motif, the catalytic groove was in the closed conformation with limited accessibility. However, in the active form, Arg220 and Arg223 were neutralized by phosphorylated Tyr216 and the catalytic cleft was in the open conformation with full substrate access. Buch *et al.* also constructed two ternary complexes of phosphorylated GSK-3β(p)-ATP binding with substrate peptide (KEEPPSPPQS(p)P) and inhibitor L803 (KEAPPAPPQS(p)P) at the substrate binding site through molecular docking, and carried out MD simulations on the two bound complexes. The conserved Phe67 from the G-loop played an important role for substrate and inhibitor binding by stabilizing the active conformation and ATP positioning. Typically, the binding modes for the substrates and inhibitors were similar with pSer strongly interacting with the positively charged triad (Arg96, Arg180, and Lys205). The analysis of RMSD (for all the Cα's of GSK-3β) and hydrogen bonds indicated a slightly tighter binding of the inhibitor 803, which was consistent with the experimental results. The observation was explained by the mutation of Glu in the substrate to Ala in the inhibitor at the P-3 position. However, a detailed binding energy analysis to explain the different binding affinities was not presented.

To better understand the primed substrate specificity, Lu *et al.* modeled the ternary complex structures of GSK-3β binding with ATP and substrate peptides with and without primed phosphorylation, and carried out MD simulations and binding free energy calculations on the constructed complex systems.[102] The ternary complexes consisting of GSK-3β with phosphorylated Tyr216, ATP, and 8-residue glycogen synthase peptides (ACE-RHSSPHQS(p)-NME and ACE-RHSSPHQS-NME) were constructed. The two complexes were referred to as p-Tyr216/GSK-3β/ATP/pGS and p-Tyr216/GSK-3β/ATP/GS for simplicity. During the MD simulations, the conformational changes of the C-loop, αC-helix, and A-loop were monitored. Although the active conformation of the A-loop of GSK-3β was observed for both primed and non-primed substrates, they had different impacts on the relative motions of the C-loop and the β-turn secondary structure of the A-loop: a closed conformation with low flexibility was induced by the primed substrate and an open conformation with high flexibility was presented in the non-primed substrate system. The priming phosphorylation at the P+4 position properly aligned the primed substrate into the active site, and the priming phosphate group stabilized the triad of positively charged residues through strong electrostatic interactions. The electrostatic potential generated by the triad could not be effectively neutralized by the non-primed substrate, and the electrostatic repulsion resulted in high flexibility of the side chains of the triad. The movement of the Arg96 side chain led to an open conformation with an enlargement of the cavity volume. Radial distribution function (RDF) analysis revealed that the water molecules around the side chains of the triad participated in the hydrogen bonding and disturbed the interactions between the triad and the non-primed substrate. Furthermore, the distance between the oxygen of the P-site serine and the γ-phosphorus of ATP (S_0-O_γ…P_γ-ATP) of the primed substrate was much shorter than that of the non-primed substrate. The shortened distance resulting from the tighter binding could facilitate the phosphate transfer reaction. MM-GBSA analysis was further performed, and large differences of binding free energies were observed for p-Tyr216/GSK-3β/ATP/pGS and p-Tyr216/GSK-3β/ATP/GS. Residue-based decomposition analysis further revealed that major favorable contributions were gained from Arg96, Arg180, Lys205, and Val214.

2.3. COMPUTATIONAL MODELING TARGETING THE C-LOBE HYDROPHOBIC GROOVE

The scaffolding peptide axin and FRAT bind competitively to GSK-3 at the same hydrophobic groove formed by the αG-helix (Gly262-Leu273) and an extended loop (Asn285-His299) near the C-lobe. The binding of the two peptides to GSK-3 is involved in the specific regulation mechanism in the Wnts signaling pathway. The substrate peptides adopted α-helical secondary structures in packing against the hydrophobic channel mainly through hydrophobic but also with a few polar interactions. The protein-substrate interface revealed a typical hydrophobic helix-helix ridge-groove interaction involving residues Val263, Leu266, Val267, and Ile270 from the αG-helix of GSK-3 (Figure 2.3) and helically disposed residues of the substrates, which are residues Leu212, Ala216, and Leu220 of FRAT and Phe388, Leu392, Leu396, and Val399 of axin. The significant difference in the binding modes of axin and FRAT involved distinct interactions with the extended loop in the C-lobe. There is a sharp turn in the FRAT peptide structure, which occurs at Gly210-Asn211 and breaks the FRAT α-helical segment into two parts. The second α-helix of FRAT occupies the same hydrophobic groove as axin does which adopts a single intact α-helix. As a result, the peptide NH groups of residues Leu212, Ile213, and Lys214 from the second α-helix of FRAT form hydrogen bond interactions with the side chains of Tyr288 and Glu290 from the extended loop in GSK-3 (Figure 2.3). However, such hydrogen bonds were not observed in the binding of axin to GSK-3.

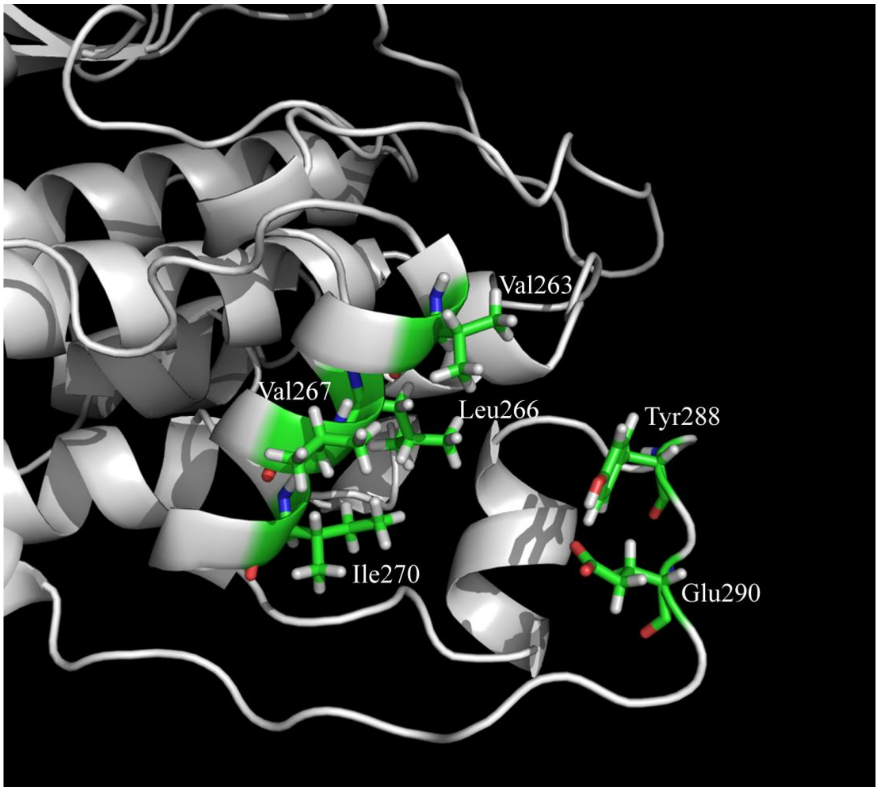


Figure 2.3. The C-lobe hydrophobic groove with important residues highlighted in stick representation (PDB code: 1GNG).

The experiments cannot fully explain the fact that different mutations of GSK-3 residues result in selective reduction in binding affinities of different substrates. Molecular modeling, especially MD simulation, provides a powerful tool to better understand the dynamic features of the structural motions and changes, so it can be incredibly helpful in terms of interpreting the experimental results of mutagenesis. Zhang *et al.* performed MD simulations on three systems: WT GSK-3β with bound axin, the V267G mutant of GSK-3β with bound axin, and GSK-3β with the bound L392P mutant of axin, followed by MM-GBSA binding energy calculations.[103] Throughout the MD simulations of different systems, GSK-3β did not undergo significant conformational changes, but the substrate axin exhibited distinct dynamical behavior. In the WT system, axin was well maintained in the hydrophobic groove. However, the mutant V267G resulted in a packing defect at the hydrophobic interface, which was demonstrated by the upward motion of axin toward αG-helix. The mutation destroyed the integrity of the hydrophobic interactions of Val267 from GSK-3β with Leu392 and Leu396 from axin, which triggered the positional shift of axin and impaired the important salt bridge interaction between Asp264 from GSK-3β and Arg395 from axin. Furthermore, the mutant L392P on axin resulted in partial helix distortion, which was observed based on the evidence of abnormal dihedral angles and decreased intra-helix hydrogen bond occupancies. The conformational distortion of axin impaired the hydrophobic interactions as well as the salt bridge between Asp264 and Arg395. The binding free energy analysis provided further evidence of the packing defect of hydrophobic interactions introduced by the mutations. Two mutations reduced significantly the van der Waals energy term, which was the dominant favorable contribution to the binding affinity for the wildtype.

The single-point mutation of Val267 to Gly on GSK-3β selectively abolished the binding affinity toward axin without impact on FRATide binding, whereas single-point mutation of Tyr288 to Phe on GSK-3β selectively abolished the FRATide binding without affecting axin binding. To provide atomic-level evidence of the mutagenesis studies, Tang *et al.* carried out MD simulations on the different GSK-3β structures bound to GSKIPtide (GSK-interacting peptide), which binds GSK-3β in a manner similar to axin.[104] The sequence alignment of three substrates (AxinGID, FRATide, and GSKIPide) revealed a common L/A-X-X-R-L motif that played an important role in protein-substrate interactions. The Leu or Ala residue at the first position participated in the hydrophobic helix-helix ridge-groove interaction, and the Arg residue formed an important salt bridge interaction with Asp264 on GSK-3β. Since GSKIPtide binds GSK-3β in a similar manner compared to AxinGID, it is not surprising that the V267G mutant of GSK-3β reduced the binding affinity of GSKIPtide by 70% and abolished the binding affinity of AxinGID. The experimental results can be explained by the observation that the reduction in volume of the hydrophobic side chain from Val to Gly distorted the hydrophobic interface and caused the positional shift of GSKIPtide residues. However, the mutation of Val267 to Gly did not affect the binding affinity of FRATide. This can be explained by the observation that Ala216 of FRATtide, which is in the corresponding position to Leu126 of GSKIPtide and Leu392 of AxinGID, contributed minimal binding affinity without any hydrophobic contact with the αG-helix of GSK-3β, because of its short side chain. Hence, the reduction in volume caused by the side chain mutation implied a packing defect for GSKIPtide and AxinGID, but not for FRATide. Another mutation of Tyr288 to Phe abolished the binding affinity for FRATide without any impact on GSKIPtide and AxinGID binding. The experimental results can be explained by the different binding modes with the extended loop on GSK-3β's C-lobe. In the complex of GSK-3β bound to GSKIPtide, the Tyr288 on GSK-3β interacted with Arg119 on the substrate in two ways: a) a hydrophobic interaction between the ring moiety of Tyr288 and the aliphatic moiety of Arg119; b) an electrostatic interaction between the hydroxyl group of Tyr288 and the positively charged guanidino group of Arg119. The mutation of Tyr288 to Phe only affected the electrostatic interaction, so it induced insignificant conformational change. In the complex of GSK-3β bound to AxinGID, the aromatic ring of Tyr288 on GSK-3β was sandwiched by the aliphatic moiety of Pro385 and the aromatic ring of Phe388 on AxinGID. So the mutation of Tyr288 to Phe did not have any impact on the hydrophobic interaction among the three residues. However, in the complex of GSK-3β bound to FRATide, the hydroxyl group of Tyr288 as well as the carboxyl group of Glu290 were strongly involved in the hydrogen bond interactions with the backbone NH group of the residues Leu212, Ala216, and Leu220 on the FRATide second α-helix. The mutation destroyed the hydrogen bonds, and subsequently impaired the FRATide binding. Electrostatic potential analysis supported the evidence that the binding modes of GSKIPtide and AxinGID are similar, but they are different from the binding mode of FRATide. The structure-based knowledge may benefit the rational design of small peptides which can selectively inhibit GSK-3 phosphorylation towards different substrates.

Chapter 3. PHARMACOPHORE MODELING, ENSEMBLE DOCKING, AND VIRTUAL SCREENING STUDIES ON GLYCOGEN SYNTHASE KINASE-3B

Gang Fu, Prasanna Sivaprakasam, Olivia R. Dale, Susan P. Manly, Stephen J. Cutler, and Robert J. Doerksen

3.1. INTRODUCTION

In the present research project, we explored the performance of implementing a mixed ligand-based/structure-based approach in the virtual screening procedure to identify structurally novel and diverse ATP-competitive GSK-3β inhibitors. For the ligand-based approach, we employed the Phase[105] module from Schrödinger 2010 to construct a common binding hypothesis for a collection of structurally diverse and highly potent GSK-3β inhibitors. The generated 3D pharmacophore model was used for preliminary screening of large databases to preclude selection of compounds lacking the key structural features necessary to be kinase inhibitors. To achieve maximum coverage of the activity space and reduce structural redundancy in the screening process, diversity analysis was used to ensure identification of a diverse set of compounds. Finally, molecular docking studies were performed to estimate the binding strengths of ligand-protein complexes. To address the protein flexibility issues, we used the ensemble docking protocol which has been shown to be superior to individual docking.

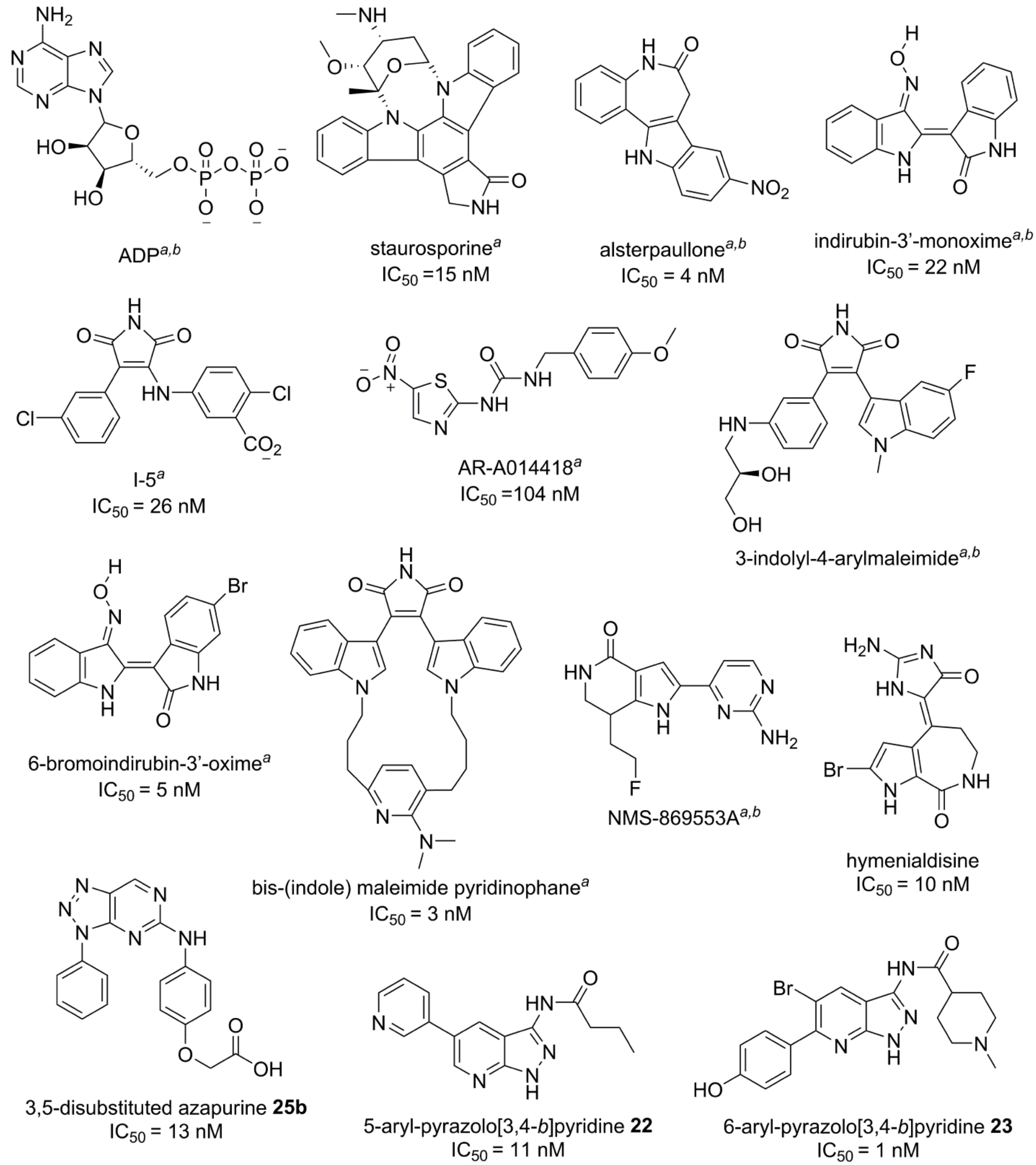
3.2. MATERIALS AND METHODS

3.2.1. PHARMACOPHORE MODELING

A collection of 22 active GSK-3β inhibitors or substrate derivatives with distinct structural features was selected for development of the 3D pharmacophore model to be used as a query to screen large databases. Figure 3.1 shows the chemical structures and the published GSK-3β inhibitory activities of those compounds. The numbering for the compounds is retained from the original publications and the references can be found in Table 3.1. The collection contains 10 compounds directly extracted from X-ray co-crystal structures with GSK-3β. In the pharmacophore model building they were restricted to have rigid conformations exactly matching the ones in the crystal structures. The rest of the compounds in the collection are highly potent and structurally distinct GSK-3β inhibitors. They have flexible conformations and so it was necessary to search for their minimum conformations before pharmacophore model generation. The MacroModel[106] module from Schrödinger 2010 was used to generate conformers. The mixed Monte Carlo multiple minimum (MCMM)/low-mode conformational search method was used, followed by 100 minimization steps with the OPLS-2005 force field. After the search the conformers which were similar within a root-mean-square deviation (RMSD) of 1.0 Å were considered as redundant conformers and eliminated. The maximum number of conformers per compound was limited to 1000. The possible pharmacophore features considered included hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic (H), negatively charged group (N), positively charged group (P), and aromatic group (R). Pharmacophore models were generated which satisfied two requirements: 1) at least three features were included in the model; 2) the model needed to match all the active compounds in the collection. Finally, the models were examined and ranked by survival score, which is calculated by the following equation:

$$\text{Survival Score} = \text{Vector Score} + \text{Site Score} + \text{Volume Score} + 1 \tag{3.1}$$

The vector score is the average of the cosines of the angles formed by corresponding pairs of A, D, or R vector features. The site score is computed based on alignment score which is the RMSD in the site-point positions. The volume score is calculated based on the overlap of van der Waals (VDWs) models of the non-hydrogen atoms in each pair of structures.



ADP^{a,b}

staurosporine^{a}
IC_{50} =15 nM

alsterpaullone^{a,b}
IC_{50} = 4 nM

indirubin-3'-monoxime^{a,b}
IC_{50} = 22 nM

I-5^{a}
IC_{50} = 26 nM

AR-A014418^{a}
IC_{50} =104 nM

3-indolyl-4-arylmaleimide^{a,b}

6-bromoindirubin-3'-oxime^{a}
IC_{50} = 5 nM

bis-(indole) maleimide pyridinophane^{a}
IC_{50} = 3 nM

NMS-869553A^{a,b}

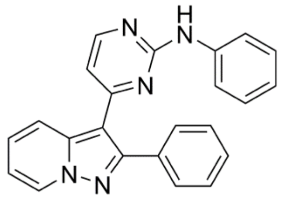
hymenialdisine
IC_{50} = 10 nM

3,5-disubstituted azapurine **25b**
IC_{50} = 13 nM

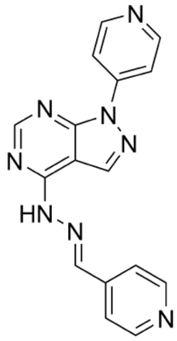
5-aryl-pyrazolo[3,4-*b*]pyridine **22**
IC_{50} = 11 nM

6-aryl-pyrazolo[3,4-*b*]pyridine **23**
IC_{50} = 1 nM

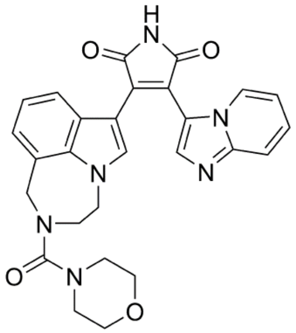
Figure 3.1. The chemical structures and biological activities of representative GSK-3β inhibitors. (a) 10 compounds directly extracted from X-ray co-crystal structures with GSK-3β; (b) compounds used only for pharmacophore modeling; (c) compounds used only for enrichment study; Compounds without superscripts are without known bioactive conformation and so underwent conformational search; they were used in both pharmacophore modeling and in the enrichment study.



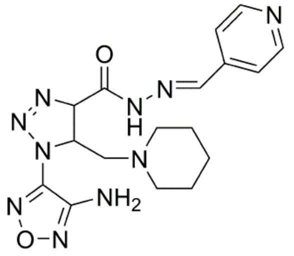
heterocycle-substituted pyrimidine **15**
IC_{50} = 60 nM



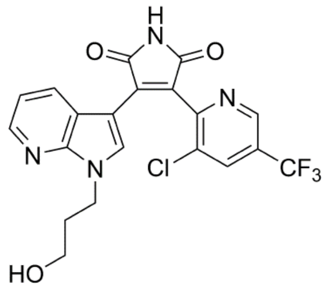
pyrazolopyrimidine **26**
IC_{50} = 8.5 nM



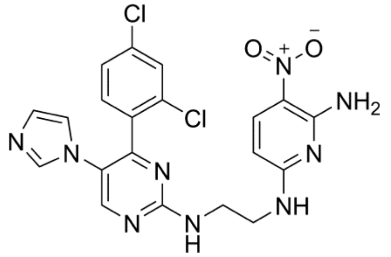
3-imidazo[1,2-*a*]pyridin-3-yl-4-(1,2,3,4-tetrahydro-[1,4]diazepino-[6,7,1-*hi*]indol-7-yl)pyrrole-2,5-dione **10**
IC_{50} = 1.3 nM



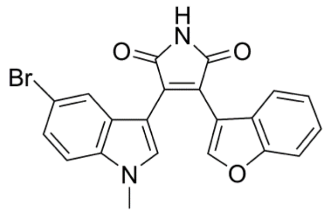
1-(4-aminofurazan-3-yl)-5-dialkylaminomethyl-1*H*-[1,2,3]triazole-4-carboxylic acid derivative **6b**
IC_{50} = 100 nM



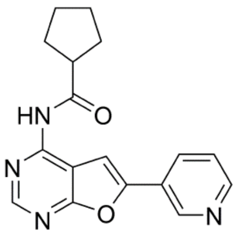
3-(7-azaindolyl)-4-arylmaleimide **17c**
IC_{50} = 26 nM



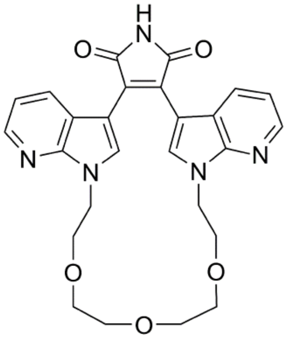
CHIR 98014
IC_{50} = 0.58 nM



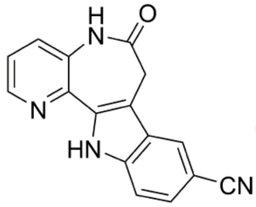
3-(benzofuran-3-yl)-4-(indol-3-yl) maleimide **2b**
IC_{50} = 7 nM



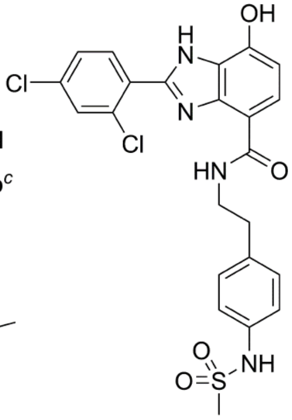
4-acylamino-6-arylfuro [2,3-*d*]pyrimidine **24**
IC_{50} = 5 nM



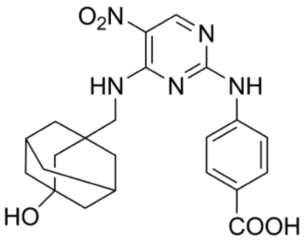
bis-7-azaindolylmaleimide **28**^{c}
IC_{50}=34nM



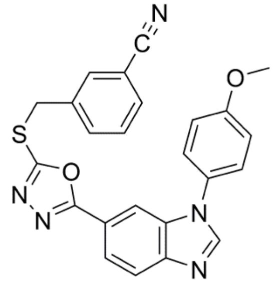
9-cyano-1-azpaullone **2b**^{c}
IC_{50} = 8 nM



7-hydroxy-1*H*-benzoimidazole derivative **6h**^{c}
IC_{50}=15nM



2,5-diaminopyrimidine **29a**^{c}
IC_{50} = 3 nM



1,3,4-oxadiazole derivative **20x**^{c}
IC_{50} = 2.3 nM

Figure 3.1. Continued.

Table 3.1. Pharmacophore modeling results including the best alignments of the chosen active compounds.

No.	Name	IC_{50} (nM)	Confs	Fitness	Relative Energy	PDB and Reference
1	ADP	n.a.	1	2.24	0	1J1C[40]
2	staurosporine[a]	15	1	2.24	0	1Q3D[41]
3	alsterpaullone	4	1	2.07	0	1Q3W[41]
4	indirubin-3'-monoxime	22	1	1.73	0	1Q41[41]
5	I-5[a]	26	1	1.65	0	1Q4L[41]
6	AR-A014418[a]	104	1	1.67	0	1Q5K[31]
7	3-indolyl-4-arylmaleimide	n.a.	1	1.72	0	1R0E[43]
8	6-bromoindirubin-3'-oxime[a]	5	1	1.75	0	1UV5[29]
9	bis-(indole) maleimide pyridinophane[a]	3	1	1.68	0	2OW3[46]
10	NMS-869553A	n.a.	1	3.00	0	3DU8[48]
11	hymenialdisine[a]	10	2	2.31	0	30
12	3,5-disubstituted azapurine **25b**[a]	13	12	1.69	0.325	107
13	5-aryl-pyrazolo[3,4-b]pyridine **22**[a]	11	4	1.91	0.127	72
14	6-aryl-pyrazolo[3,4-b]pyridine **23**[a]	1	3	1.78	0	70
15	heterocycle-substituted pyrimidine **15**[a]	60	13	1.27	0.315	108
16	pyrazolopyrimidine **26**[a]	8.5	2	2.32	0	109
17	3-imidazo[1,2-a]pyridin-3-yl-4-(1,2,3,4-tetrahydro-[1,4]diazepino-[6,7,1-hi]indol-7-yl)pyrrole-2,5-dione **10**[a]	1.3	31	1.77	2.597	110
18	1-(4-aminofurazan-3-yl)-5-dialkylaminomethyl-1H-[1,2,3]triazole-4-carboxylic acid derivative **6b**[a]	100	66	2.17	2.400	111
19	4-acylamino-6-arylfuro[2,3-	5	8	2.14	3.834	112

	d]pyrimidine **24**[a]					
20	CHIR 98014[a]	0.58	68	1.58	1.086	113
21	3-(benzofuran-3-yl)-4-(indol-3-yl)maleimide **2b**[a]	7	8	1.77	0.934	114
22	3-(7-azaindolyl)-4-arylmaleimide **17c**[a]	26	71	1.82	3.495	115

[a] Compounds used for both pharmacophore modeling and enrichment study.

3.2.2. DIVERSITY ANALYSIS

Structurally similar compounds are likely to exhibit similar activity, and hence maximum coverage of the activity space should be achieved by selecting a structurally diverse set of compounds. To measure shape similarity, linear hashed binary fingerprints were calculated. This has been shown to be an efficient approach to cluster large data sets and evaluate the diversity of compound libraries.[116] Since the database to be screened is large (>400,000 compounds), the shape fingerprints were calculated at 64-bit precision with maximum linear path equaling 14 in order to be able to distinguish molecules in the database to a large extent. After the shape fingerprints calculations, dissimilarity-based compound selection (DBCS) was performed, which was based on the calculated pairwise Soergel distances between compounds. The sum of the distances between each compound and the selected subset was evaluated at each round and the compound with the largest total distance from the selected subset was added. This method is called maximum sum of distances and is initialized by selecting representative compounds. This is the most efficient and effective non-hierarchical clustering method.[117] After diversity analysis, about 8,000 structurally distinct compounds were selected to conduct docking screening. The shape-fingerprint calculations and DBCS analysis were performed using Canvas[118] implemented in Schrödinger 2010.

3.2.3. DOCKING STUDIES

The structural features of ligand-protein interactions provide insight into the important binding features which can be useful for discovering and designing novel potent and selective inhibitors for GSK-3β. To estimate the reliability and prediction accuracy of various docking protocols, several validation experiments have been conducted including self-docking and cross-docking. In self-docking, the native ligand (from a protein-ligand co-crystallized structure) is docked back into its corresponding receptor structure, whereas in cross-docking every extracted ligand is docked into all the receptor structures in the collection. In the present study, a total of 20 crystal structures have been culled for GSK-3β either with or without ligands in the ATP binding pocket. The list of all the crystal structures with their Protein Data Bank (PDB) code, ligands and references are tabulated in Table 3.2. If the crystal structure contained more than one GSK-3β chain, then only chain A was considered. The *apo* PDB structures 1GNG, 1H8F, and 1I09 were not included in the following docking investigations since they do not have important water molecules in the binding sites, and their local conformational features around the binding pockets are different from the ligand-bound protein structures especially in the glycine rich loop and hinge region because of the absence of induced fit effects. Since 1J1B and 1PYX have the same ligands in the binding pocket, only 1PYX was included in the following studies. The PDB structure 2JLD demonstrates the binding of GSK-3β to a ruthenium complex but the corresponding force field parameters for ruthenium are not available in the docking programs. The PDB structure 3F88 contains a ligand missing important connection bonds. So those two crystal structures were excluded in the self-docking and cross-docking experiments, but they were included in the enrichment studies since the receptor structures had undergone induced fit effects to accommodate their native ligands. The PBD structures 1J1C, 1O9U, and 1PYX have as native ligands ADP, ADZ, and ANP, respectively. These are ATP derivatives and have highly negatively charged functional groups which are surrounded by a number of water molecules in the crystal structures, so they have different induced fit compared to the structures containing small molecule inhibitors. Hence those three crystal structures were not included in the cross-docking experiments. To evaluate the prediction accuracy, the RMSD between the docking poses and experimental structures of the ligands were calculated, and the lower RMSD indicates the better docked pose. In order to determine which docking algorithm is best for these proteins, several leading docking programs were examined and compared, including Glide 5.6,[119] GOLD 4.0,[120-122] AutoDock 4.2,[123, 124] and MOE 2010.10 Dock.[125] Then ensemble docking was also carried out in contrast to individual docking to probe sensitivity to induced fit effects during ligand-protein recognition.

Table 3.2. Available GSK-3β X-ray crystal structures from PDB

PDB code	Substrate	Resolution[b]	Release Date	Reference
1GNG	none	2.60	2002-10-03	11
1H8F	none	2.80	2002-01-31	7
1I09	none	2.70	2002-01-01	8
1J1B[a]	ANP	1.80	2003-12-03	40
1J1C[a]	ADP	2.10	2003-12-03	40
1O9U	ADZ	2.40	2003-08-15	10
1PYX[a]	ANP	2.40	2003-10-21	41
1Q3D[a]	Staurosporine	2.20	2003-10-21	41
1Q3W[a]	Alsterpaullone	2.30	2003-10-21	41
1Q41[a]	Indirubin-3'-monoxime	2.10	2003-10-21	41
1Q4L[a]	I-5	2.77	2003-10-14	41
1Q5K[a]	AR-A014418	1.94	2004-08-10	31

1R0E[a]	3-indolyl-4-arylmaleimide	2.25	2004-10-12	43
1UV5	6-bromoindirubin-3’-oxime	2.80	2004-01-29	29
2O5K	7-hydroxy-1H-benzoimidazole	3.20	2007-10-23	45
2OW3[a]	bis-(indole)maleimide pyridinophane	2.80	2008-02-19	46
2JLD[a]	Ruthenium complex	2.35	2008-12-09	126
3DU8[a]	NMS-869553A	2.20	2009-03-03	48
3F7Z[a]	1,3,4-oxadiazole	2.40	2009-03-10	127
3F88[a]	1,3,4-oxadiazole	2.60	2009-03-10	127

[a] Only chain A was considered. [b] In Å.

Glide 5.6. Before performing the docking experiments, the receptor crystal structures went through protein structure preparation, which included deleting crystallographic water molecules beyond 5 Å from the native ligands, adding hydrogens, assigning bond orders and ionization states, optimizing hydrogen bond networks, and finally minimizing protein structures. All the preparation jobs were conducted using the Protein Preparation Wizard implemented in Schrödinger 2010 with default setup. The ligand structures were prepared using the LigPrep module. The ionization states and stereochemistry of the ligands were maintained to be the same as reported. The docking program Glide (Grid-based Ligand Docking with Energetics) uses a series of hierarchical filters to determine the possible locations of the ligand in the binding site of the receptor. The hierarchy includes an initial exhaustive search of all possible conformations of ligands followed by greedy scoring and refinement using the Schrödinger discretized version of the ChemScore empirical scoring function, subsequently followed by grid minimization and final scoring using GlideScore. A grid which represents the shape and properties of the receptor by various sets of VDWs and electrostatic fields is used to align and score the ligand poses in the binding site. To generate the scoring grids, the default VDW radius scaling factor (1.0) and partial charge cutoff (0.25) were used. The grids were centered around the centroid of native ligands in the structures, with enclosing box (outer box) dimensions of 34 Å × 34 Å × 34 Å and ligand diameter midpoint box (inner box) dimensions of 14 Å × 14 Å × 14 Å. No constraints were imposed on any of the receptor structures. To perform docking experiments, the VDW radius scaling factor (0.8) and partial charge cutoff (0.15) were kept unchanged. There are three options of docking precision, which are HTVS (high-throughput virtual screening), SP (standard precision), and XP (extra precision). The Glide HTVS is fast and intended for the rapid screening of very large databases, but it has restricted conformational sampling which can result in an unsatisfactory docking pose. The Glide SP is appropriate for screening ligands in large numbers and is the default option. The Glide XP scoring function is designed to identify ligand poses that are supposed to have unfavorable energies, which is useful to exclude false positives and to provide a better correlation between good poses and good scores. It includes additional terms relative to the SP scoring function to deal with hydrophobic interactions and is designed for use only on good ligand poses. It is a powerful and discriminating procedure, but takes much more computational effort.

GOLD 4.0. GOLD (Genetic Optimisation for Ligand Docking) is a genetic algorithm for docking flexible ligands into the binding pockets of receptors. The structures for both receptors and ligands were prepared using the same procedures as were used in Glide docking. Since water molecules play key roles in ligand-receptor recognition, all the crystallographic water molecules remaining in the receptor structures were switched on and the orientation of the hydrogen atoms of active water molecules were optimized automatically using the GOLD software. To reward water displacement in the binding pocket, an additional parameter (Sbar) is added to the fitness function for both GoldScore (GS) and ChemScore (CS), except for in the Astex Scoring

Potential (ASP). The number of genetic algorithm runs was set to 10 and the default early termination option was turned on to stop docking runs as soon as the top three solutions were within the RMSD values of 1.5 Å. To define the binding site, all protein atoms within 10 Å of the native ligand and their associated residues were considered. The LIGSITE cavity detection algorithm was used to restrict the region of interest to the solvent accessible surfaces. Finally, the docking analyses were carried out for all three fitness functions, GS, CS, and ASP. The GS fitness function has been optimized for the prediction of ligand binding poses, so it can be used to discriminate between different binding modes of the same ligand molecules. However, obtaining a significant correlation between a fitness score and the biological activities is not guaranteed. Unlike the GS fitness function which is based on force field parameters, the CS fitness function is derived empirically from a regression model based on a set of 82 protein-ligand complexes for which biological activities are available. In contract to the GS and CS fitness function, ASP is an atom-atom potential derived from the statistical potentials generated by analyzing existing ligand-protein structures in the entire PDB database.

AutoDock 4.2. AutoDockTools implemented in MGLTools 1.5.4 developed at the Molecular Graphics Lab (MGL) of the Scripps Research Institute were used to set up, run and analyze AutoDock 4.2 dockings. To prepare the receptor structures, hydrogens were added, Kollman United atomic charges were assigned, and non-polar hydrogens were merged with their parent carbon atoms. To prepare the ligand structures, partial atomic charges were assigned using the Gasteiger method after all hydrogens were added and torsion trees were set up for the rotational bonds using AutoTors. Grid boxes of 40 points in each dimension with spacing of 0.375 Å between grid points were constructed for all the receptors using AutoGrid, centering on the centroid of the native ligands. To perform docking studies, three search methods were used including the Solis & Wets algorithm for local search, the Genetic Algorithm (GA) for global search, and the Lamarckian GA (LGA) for hybrid global-local search. The maximum iterations were set to 300 for the Solis & Wets local search. The number of individuals in the population of 150, the maximum number of generations of 27,000 and the maximum number of energy evaluations of 2,500,000 were set up for the GA global search. Ten runs of hybrid LGA search were performed with the local search frequency set to 0.06. Finally energy-ranked cluster analyses were carried out on the docking results with the RMSD cluster tolerance of 2.0 Å. The free-energy scoring function of AutoDock 4.2 is based on linear regression analysis of a large set of diverse protein-ligand complexes with known biological activities using the AMBER force field.

MOE 2010.10 Dock. To add hydrogens to the receptor and ligand structures, Protonate 3D was used with default settings. Then partial charges for both receptors and ligands were calculated using the MMFF94 force field[128]. To generate conformations for the rotatable bonds, the Triangle Matcher placement method was used. It works by systematically superimposing ligand atom triplets to the triplets of receptor site points. The receptor site points are alpha sphere centers which represent locations of tight packing. The default London dG (LdG) scoring function was used to estimate the free energy of binding of the ligand from a given pose. The top ten ranked docking poses were retained for further refinement. Two refinement schemes were carried out after scoring. One is ForceField refinement (FFR) by which energy minimization of the system is performed using the current forcefield (MMFF94) and the final energy is evaluated using the Generalized Born solvation model (GB/VI); the other one is GridMin refinement (GMR) in which a grid is used for electrostatic calculations during the minimization process which can speed up the process. The VDW interactions are treated explicitly and the distance-dependent dielectric model is used. The final electrostatic energy is calculated using the explicit Coulomb form instead of the grid.

Ensemble Docking. An ensemble of 13 protein structures was used for docking studies, including 1Q3D, 1Q3W, 1Q4L, 1Q41, 1Q5K, 1R0E, 1UV5, 2JLD, 2O5K, 2OW3, 3DU8, 3F7Z, and 3F88. For each ligand, the top-ranked pose against each receptor structure was saved, and the ensemble of docking poses for each ligand was sorted by GlideScore. Since for some ligands we could not obtain a reasonable docking pose against certain receptor structures using Glide, the number of members of the ensemble of the top-ranked docking poses for each ligand is less than or equal to 13. After ranking, only the pose with the lowest GlideScore was retained for each ligand, and the other poses were eliminated. The post-docking processes, including pooling, sorting, and redundancy elimination, were conducted using the utility program ‘glide_merge’ implemented by Schrödinger. This docking protocol is called ensemble docking. In contrast to individual docking in which only a single receptor structure is used, ensemble docking takes into account the protein flexibility to some extent. The performance in terms of the docking predictions of both individual docking and ensemble docking was evaluated using enrichment studies.

3.2.4. ENRICHMENT STUDIES

A collection of 22 highly active GSK-3β inhibitors with distinct structures was selected and seeded into a 1,000 decoy dataset created by Schrödinger.[129, 130] The random hit rate was 2.2%, which means without any assistance of predictive algorithm, the compound randomly selected from the combined dataset has a 2.2% chance to be an active inhibitor. The collection of highly active GSK-3β inhibitors (Figure 3.1) includes 17 compounds used in the pharmacophore modeling and 5 other compounds, which are 7-hydroxy-1H-benzoimidazole derivative **6h**,[45] bis-

7-azaindolylmaleimide **28**,[131] 1,3,4-oxadiazole derivative **20x**,[127] 9-cyano-1-azapaullone **2b**,[28] and 2,5-diaminopyrimidine **29a**.[107] The decoy dataset was created by selecting ligands from a library containing one million compounds that exhibit 'drug-like' properties. To evaluate the effectiveness of the different virtual screening protocols using individual docking and ensemble docking, the hit rates (Hits%) and enrichment factors (EF) of the known active compounds within 1, 5, and 10% of the top-ranked compounds were calculated. The hit rate is calculated using the following equation:

$$\text{Hits\%} = \left(\frac{\text{Hits}_{\text{active}}}{\text{Total}_{\text{active}}}\right) \times 100 \qquad (3.2)$$

$\text{Hits}_{\text{active}}$ is the number of known active compounds in the top-ranked list, and $\text{Total}_{\text{active}}$ is the number of total known active compounds, which is 22. The enrichment factor is represented by:

$$\text{EF} = \frac{\text{Hits}_{\text{active}} \,/\, \text{Total}_{\text{active}}}{\text{Hits}_{\text{decoy}} \,/\, \text{Total}_{\text{decoy}}} \qquad (3.3)$$

$\text{Hits}_{\text{decoy}}$ is the number of decoy compounds in the top-ranked list, and $\text{Total}_{\text{decoy}}$ is the number of compounds in the total decoy dataset, which is 1,000. The enrichment curves were also obtained for both individual docking and ensemble docking. The Hits% and EF were calculated based on the assumption that all the compounds in the decoy dataset are inactive against GSK-3β, however, that cannot be guaranteed. There may be some compounds in the decoy dataset that actually are active against the target protein. Thus, an enrichment study is just an estimate of the screening effectiveness.

3.2.5. HIGH THROUGHPUT VIRTUAL SCREENING

The ChemBridge EXPRESS-Pick$^{\text{TM}}$ database consisting of more than 480,000 small synthetic compounds was selected for performing virtual screening, which is illustrated in Figure 3.2. The 3D structures of those compounds were generated using the LigPrep[132] module implemented in Schrödinger 2010. To filter the compounds for drug-likeness, the modified Lipinski's rules were used, which are as follows: molecular weights are in the range of 200 to 600; number of hydrogen bond donors is from 0 to 6; number of hydrogen bond acceptors is from 0 to 12; calculated logP values are from 1 to 5; and the number of rotatable bonds is from 0 to 15. A set of 441,612 compounds survived into the next stage, which was screened using the common pharmacophore hypothesis. The hypothesis filtered out about half of the screened compounds. The remaining compounds underwent diversity analysis which reduced the number of compounds to ~12,000. Those compounds were subjected to the last stage of the virtual screening procedure, ensemble docking using the 13 protein crystal structures mentioned above. The hit compounds with preferable Glide SP scores and reasonable docking poses were selected and subjected to the enzyme inhibition assays.

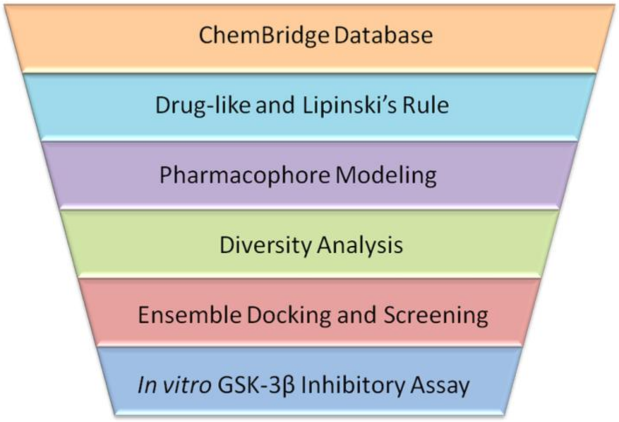

ChemBridge Database
Drug-like and Lipinski's Rule
Pharmacophore Modeling
Diversity Analysis
Ensemble Docking and Screening
In vitro GSK-3β Inhibitory Assay

Figure 3.2. Flowchart for high throughput virtual screening.

3.2.6. BIOLOGICAL VALIDATION

The Invitrogen Z-Lyte kit was used to screen for potential GSK-3β inhibitors through fluorescence resonance energy transfer (FRET) between the donor, coumarin, and the acceptor, fluorescein. The donor and acceptor are on each end of the peptide substrate, constituting the

FRET pair. The Z'-LyteTM Ser/Thr 9 peptide is used since it is designed with a priming phosphate which satisfies the substrate specificity of GSK-3β. The inhibitors were screened in a two-step reaction. In the first step, the kinase reaction proceeds, in which the γ-phosphate of ATP is transferred to a single serine or threonine residue in the synthetic peptide substrate. In the second step, the development reaction occurs, in which a site-specific protease recognizes and cleaves non-phosphorylated peptides. The cleavage disrupts FRET between the donor and acceptor fluorophores on the peptide, giving off a high emission ratio. The phosphorylated peptides remain uncleaved and maintain FRET, giving off a low emission ratio. Therefore, the kinase inhibition can be recognized by a high emission ratio which means the low phosphorylation by GSK-3β. This ratiometric approach reduces the effects of well-to-well variation, and provides a quick and reliable approach for the assessment of GSK-3β inhibition.

The enzyme inhibition assays were performed in 384-well Greiner Bio-One flat bottom microplates. First, three control solutions were prepared. The maximum emission ratio was established based on the 100% inhibition control, which yields 100% cleaved peptide in the development reaction. The 100% inhibition control contained 2.5 μL of 4% DMSO, 5 μL of kinase/peptide mixture, and 2.5 μL of diluted kinase buffer. The original buffer, which was prepared by mixing 250 mM HEPES, 50 mM $MgCl_2$, 5 mM EGTA, and 0.05% BRIJ-35 at pH 7.5, was diluted by 3.76-fold. The minimum emission ratio was established as the 0% inhibition control, which is designed to produce a recommended 20-40% phosphorylated peptide in the kinase reaction and to yield 60-80% cleaved peptide in the development reaction. The 0% inhibition control contained 2.5 μL of 40 μM ATP instead of 2.5 μL of kinase buffer, resulting in a 10 μM final ATP concentration. The 100% phosphorylation control, consisting of 5 μL synthetically phosphorylated peptide instead of 5 μL of kinase/peptide mixture, is designed for calculation of percent phosphorylation. This control yields a very low percentage of cleaved peptide in the development reaction. The inhibitory activities of the hit compounds against GSK-3β were assayed by reading the fluorescence signals of the reaction mixture, which contained 2.5 μL of 40 μM hit compound dissolved in 4% DMSO, 5 μL of kinase/peptide mixture, and 2.5 μL of 40 μM ATP. The final concentration of each compound to be tested in the assay was 10 μM, and the final concentration of DMSO never exceeded 1%. The percent inhibition is calculated by the following equation:

$$\%\text{Inhibition} = 100 \times \left(1 - \frac{\%\text{phosphorylation of hit compounds}}{\%\text{phosphorylation of } 0\% \text{ inhibition control}}\right) \tag{3.4}$$

The percent phosphorylation can be calculated based on the emission ratio.

To determine the IC_{50} of the hit compounds with percent inhibition greater than 50% at 10 μM concentration, 10 concentration points were prepared using 3-fold dilutions, ranging from 10 μM to 0.5 nM, and were used for the nonlinear regression analysis of the dose response curve. For a proper comparison, indirubin-3’-monoxime (IC_{50} = 190 nM[133]) was employed as the reference standard.

3.3. RESULTS AND DISCUSSION

3.3.1. PHARMACOPHORE MODELING

After clustering and scoring, three pharmacophore models were generated, all of which have the same three features (‘A’, ‘D’, and ‘R’) with different relative positions. The pharmacophore models and their scores are listed in Table 3.3. The model with the highest survival score (2.87) was selected as the pharmacophore model to be used for screening of large molecular databases, for which the 3D coordinates and inter-feature distances and angles are listed in the Tables 3.4 and 3.5. The pharmacophore features ‘A’ and ‘D’ represent the hydrogen bond interactions with the hinge region and co-crystallized water molecule within the ATP-binding pocket. The backbone amides of the hinge region residues, including Tyr134, Val135, and Pro136, strongly interact with the ligands, and this interaction plays a key role in the ligand-protein interaction at the ATP-binding pocket.[134] The planar aromatic group (‘R’) is buried into the adenine-binding domain and interacts with the protein through hydrophobic interactions. The superimpositions of the pharmacophore model with representative compounds are shown in Figure 3.3. The best alignments of the selected active compounds and their inhibitory activities are tabulated in Table 3.1. Although the correlations between the biological activities and fitness values are low, this common pharmacophore model matches all of the diverse GSK-3β inhibitors, so it is suitable to be used as the pre-filter in database screening.

Table 3.3. Generated pharmacophore models.

No.	Features	Survival Score	Vector Score	Site Score	Volume Score	Matches
1	ADH	2.870	0.859	0.660	0.348	22
2	ADH	2.738	0.817	0.590	0.333	22
3	ADH	2.435	0.730	0.410	0.297	22

Table 3.4. The xyz coordinates for the best feature-based pharmacophore model, which was used for virtual screening.

	X[a]	Y[a]	Z[a]
A	24.313	−14.622	−9.8282
D	23.569	−13.511	−11.878
R	21.293	−16.493	−9.5971

[a] In Å.

Table 3.5. The inter-feature distances and angles for the best feature-based pharmacophore model, which was used for virtual screening.

Features	Distance (Å)	Features	Angles (°)
A—D	2.447	D—A—R	92.0
A—R	3.560	A—D—R	54.1
D—R	4.391	A—R—D	33.8

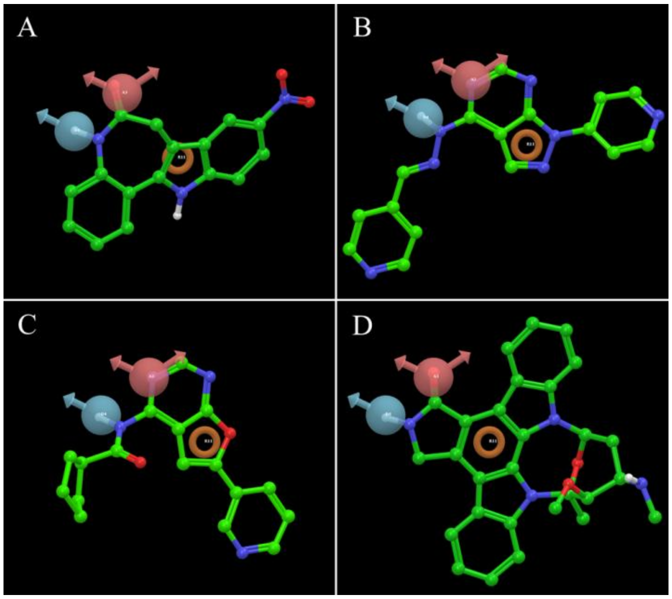


Figure 3.3. Overlap of representative GSK-3β inhibitors with the pharmacophore model having the highest survival score. A. alsterpaullone; B. pyrazolopyrimidine 26; C. 4-acylamino-6-arylfuro[2,3-d]pyrimidine 24; D. staurosporine. The pharmacophore features 'ADR' (acceptor, donor, and aromatic) are represented by the blue ball with arrow, red ball with two arrows, and orange ring, respectively. The color codes for the inhibitors are as follows: green (carbon), red (oxygen), blue (nitrogen), and white (hydrogen).

3.3.2. DOCKING VALIDATION

The RMSD values which show the displacement of docking poses relative to the experimental structures are listed in Table 3.6. For Glide 5.6, the prediction accuracies were improved with increased accuracy of docking scoring function from HTVS (high-throughput virtual screening) to SP (standard precision) to XP (extra precision). However, the computational effort increased at the same time. It is clear that Glide SP and Glide XP outperformed other docking protocols as measured by lower RMSD values. For Gold 4.0, the GoldScore (GS) scoring function gave the best performance. Some of the docking poses generated by Gold were far from the experimental structures. So using default parameters, the heuristic genetic algorithm did not identify the correct docking pose in an efficient manner. AutoDock 4 produced very good docking poses using the global search and hybrid search methods. For MOE 2010.10 Dock, the refinement processes after scoring significantly boosted the docking performance. The means and medians were obtained for all of the docking protocols, and we found that Glide XP gave the lowest values for both mean (0.84 Å) and median (0.37 Å). Glide SP was the second best protocol, with mean and median of RMSD values of 0.89 Å and 0.44 Å, respectively. Glide XP requires significantly increased computational effort compared to Glide SP. In the context of virtual screening, not only the accuracy but also the efficiency needs to be considered. Thus, Glide SP was selected as the optimal docking protocol for performing further studies including virtual screening.

Table 3.6. RMSD values for docking validation studies.

	RMSD (All atoms)											
PDB Code	Glide			Gold			MOE			AutoDock		
	HTVS	SP	XP	GS	CS	ASP	LdG	FFR	GMR	LGA	GA	LS
1J1C	1.27	1.01	1.58	1.12	1.19	0.85	3.63	3.69	3.07	0.51	2.35	1.20
1O9U	15.94	3.43	3.45	4.80	4.69	4.76	3.11	0.65	3.79	3.97	3.96	4.42
1PYX	12.19	1.69	1.70	1.37	1.01	1.42	2.42	0.89	3.03	0.93	1.00	6.17
1Q3D	0.21	0.19	0.15	0.35	0.49	0.48	1.62	0.17	0.88	0.38	0.32	1.15
1Q3W	0.08	0.07	0.13	0.61	0.27	0.21	7.08	0.16	0.44	0.25	0.22	0.90
1Q41	2.93	0.55	0.36	0.68	0.59	0.42	1.12	0.42	0.85	0.43	0.44	2.04
1Q4L	0.90	2.04	0.13	1.13	0.51	0.53	6.93	1.39	2.86	0.43	0.41	3.19
1Q5K	0.14	0.14	0.51	0.74	9.06	0.41	7.10	1.44	1.19	0.46	0.41	1.40
1R0E	6.19	0.08	0.10	1.60	6.55	6.36	6.01	6.68	6.14	0.87	2.95	3.49
1UV5	0.23	0.30	0.34	0.56	0.71	0.25	6.42	0.45	0.37	0.48	0.59	3.40
2O5K	4.36	1.74	1.73	2.91	2.18	1.18	4.25	3.45	2.60	2.33	2.00	3.89
2OW3	0.15	0.54	0.38	0.45	0.72	0.74	1.01	0.59	0.76	0.59	0.61	0.61
3DU8	6.87	0.33	0.26	0.53	0.68	0.67	3.55	0.28	0.85	0.37	0.38	5.00
3F7Z	0.91	0.28	0.94	4.63	5.72	5.80	7.16	3.58	4.25	4.15	0.73	3.45
Mean	3.74	0.89	0.84	1.53	2.46	1.72	4.39	1.70	2.22	1.15	1.17	2.88
Median	1.09	0.44	0.37	0.93	0.87	0.71	3.94	0.77	1.90	0.50	0.60	3.30

Color Code	RMSD < 1.5	1.5 < RMSD < 3.0	3.0 < RMSD < 4.5	4.5 < RMSD

To study the induced fit effects on ligand-protein binding, the significance of the co-crystallized water molecules within the ATP-binding pocket was examined (Figure 3.4). In the crystal structure 1Q3W, there is one water molecule interacting with both alsterpaullone and residue Tyr134 in the hinge region. In the crystal structure 1R0E, a water molecule located in another position plays an important role in the connection of 3-indolyl-4-arylmaleimide with residues Glu97 and Asp200. In the crystal structure 1Q3D, a hydrogen bonding network consisting of three water molecules bridges the interaction between staurosporine and residues Val135, Glu137, and Gln185, at the bottom of the binding pocket. In the crystal structure 1UV5, a distinct hydrogen bonding network consisting of three water molecules bridges the interaction between 6-bromoindirubin-3'-oxime and residues Thr138, Arg141, and Gln185. So the co-crystallized water molecules indeed play key roles in the ligand-protein interactions, and due to the induced fit effects the ligands with distinct structures may have different numbers of and locations of water molecules to enhance ligand-protein binding.

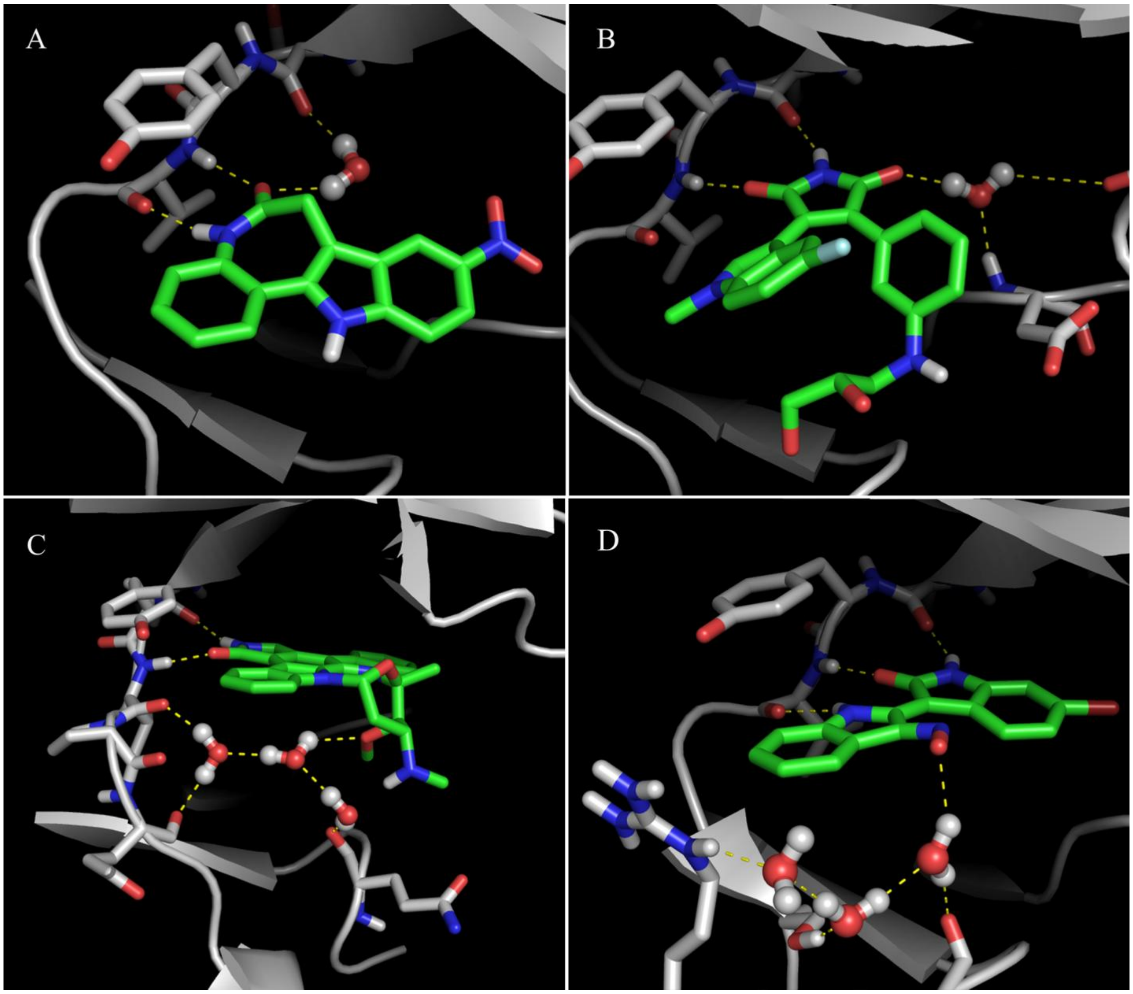


Figure 3.4. Different locations and interactions of co-crystallized water molecules within the ATP-binding pocket. A. 1Q3W; B. 1R0E; C. 1Q3D; D. 1UV5. The proteins are represented by cartoons, while the residues on the hinge region are represented by sticks. The carbons of the proteins are colored grey, while the carbons of the inhibitors are highlighted in green. The other atoms have the color codes: red (oxygen), blue (nitrogen), and white (hydrogen).

The conformational flexibility of the residues around the ATP-binding pocket can also be seen by overlapping five crystal structures, including 1Q41, 1Q4L, 1UV5, 2O5K, and 2OW3 (Figure 3.5). The co-crystallized ligands were removed before alignment for clear comparison, and the superimposition was based on the alignment of the backbone amino-acids. From the overlap, it is apparent that only Tyr134 in the hinge region has relatively little flexibility, and the side-chains of the other residues, Arg141, Glu185, and Phe67, are highly flexible. Those residues play important roles in the ligand-protein binding through hydrogen bonds and hydrophobic interactions.

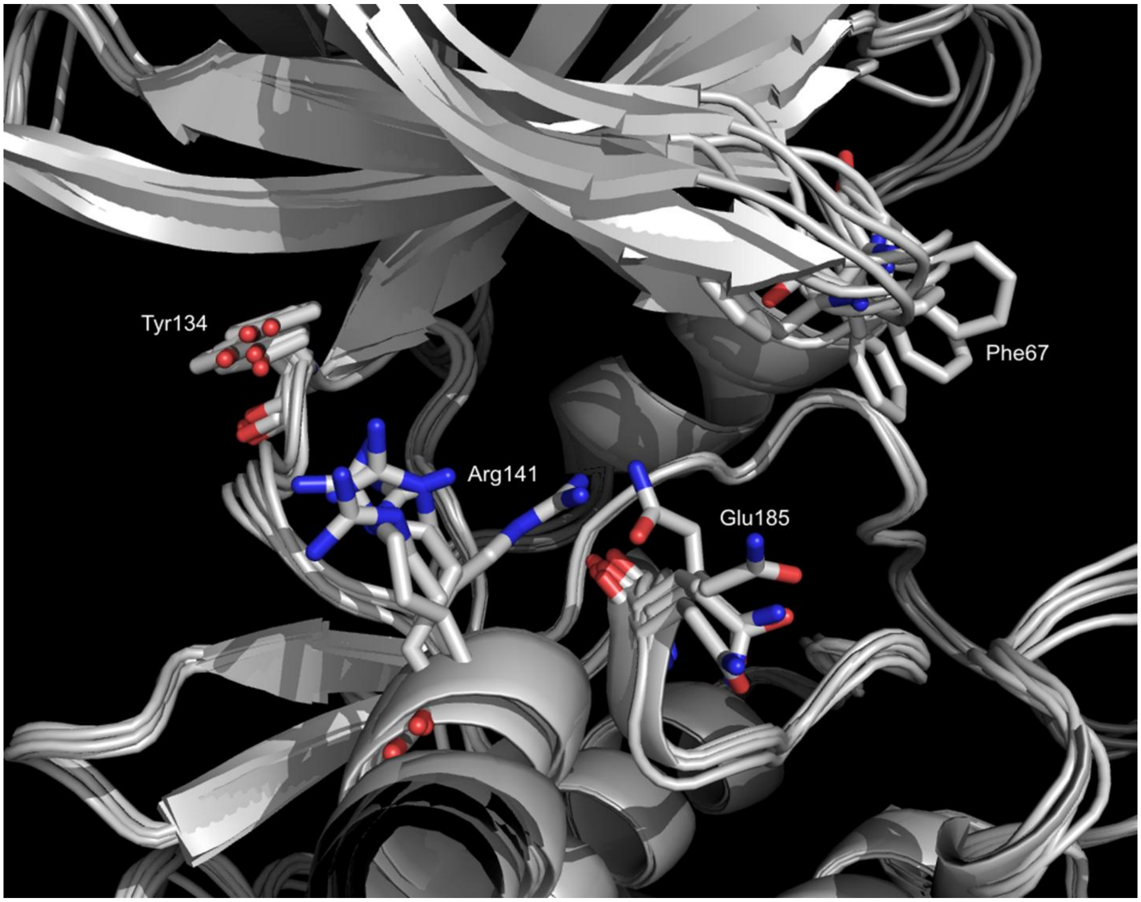


Figure 3.5. Overlap of five crystal structures, including 1Q41, 1Q4L, 1UV5, 2O5K, and 2OW3. The proteins are represented by cartoons. The residues Tyr134, Arg141, Glu185, and Phe67 are represented by sticks. The color codes are as follows: grey (carbon), red (oxygen), and blue (nitrogen).

To assess the induced fit effects on the docking disposition, self-docking and cross-docking studies were conducted using Glide SP docking. The compound bis-(indole)maleimide pyridinophane from the crystal structure 2OW3 could not be docked into the receptor in the crystal structures 1Q4L, 1Q41, and 1UV5 with a reasonable docking pose. The RMSD values of the self-docking and cross-docking of the other compounds were tabulated in Table 3.7. It is clear that self-docking yielded proper binding poses similar to the ones in the crystal structures with average RMSD equal to 0.57 Å, but the cross-docking produced binding poses far from the ones in the crystal structures with average RMSD equal to 5.04 Å. Therefore, the induced fit effects have significant influence on the docking disposition, and docking into a single crystal structure alone is not sufficient to be used to predict the binding poses of structurally diverse compounds.

Table 3.7. RMSD values for self-docking and cross-docking studies. Each row represents the ligand and each column represents the protein.

	1Q3D	1Q3W	1Q4L	1Q5K	1Q41	1R0E	1UV5	2O5K	2OW3	3DU8	3F7Z
1Q3D	0.19	5.56	3.93	10.17	8.40	7.20	9.54	4.57	6.55	4.23	6.37
1Q3W	2.05	0.07	1.27	8.66	6.99	0.91	1.78	5.22	1.20	1.06	0.95
1Q4L	7.78	3.74	0.55	8.60	3.09	2.01	3.34	3.91	4.26	5.78	5.01
1Q5K	2.46	2.83	9.72	2.04	2.03	1.35	2.08	8.69	9.25	9.02	8.41
1Q41	1.78	6.84	1.12	10.71	0.14	1.19	0.39	1.07	1.68	6.16	6.50
1R0E	6.37	5.76	3.55	7.21	6.53	0.08	3.74	6.04	1.32	5.70	5.89
1UV5	1.77	2.73	1.17	9.68	0.31	7.14	0.30	6.43	1.93	1.83	6.91
2O5K	4.49	6.96	11.64	12.23	3.92	2.59	2.36	1.74	2.99	3.36	2.62
2OW3	7.28	8.61	n.a.[a]	10.02	n.a.[a]	5.58	n.a.[a]	10.00	0.54	7.51	7.81
3DU8	1.39	6.72	6.81	3.53	5.85	5.79	6.00	6.17	2.00	0.33	0.34
3F7Z	4.33	4.57	7.65	9.25	6.35	5.88	5.89	1.64	7.21	2.12	0.28

[a] No reasonable docking pose was found.

To address the induced fit effects, the ensemble docking approach described in the methods section was used. In contrast to the induced fit docking approach[135] implemented in Schrödinger 2010, the ensemble docking method does not require protein sampling, which costs huge computational effort and is inappropriate for high throughput virtual screening. Meanwhile, ensemble docking takes into account both interactions with the water molecules and the conformational flexibility of the residues, based on evidence from the experimental crystal structures. In order to validate that ensemble docking is superior to individual docking and is appropriate to address the induced fit effects, the enrichment study described in the methods section was conducted. The performance of individual docking and ensemble docking in retrieving known active compounds from large decoy datasets was evaluated and the results are listed in Table 3.8. It is clear that only one individual docking protocol, based on crystal structure 2OW3, slightly outperformed ensemble docking, and that was only for screening of 1% of the compounds; otherwise ensemble docking significantly outperformed all the individual dockings at both 5% and 10% of the screened compounds. The enrichment curve in Figure 3.6 also shows that the ensemble docking method has the best enrichment among all the docking protocols, especially from 5% to 45% of the screened database. The cut-off line for virtual screening usually lies within this range. Thus, based on the enrichment study, ensemble docking is believed to be the most effective and accurate docking method and was used in the last stage of the virtual screening procedure.

Table 3.8. Enrichment Study.

	1%			5%			10%		
Protein	No.	Hits%	EF	No.	Hits%	EF	No.	Hits%	EF
1Q3D	5	23	23.2	11	50	10.0	12	55	5.5
1Q3W	4	18	18.6	5	23	4.6	6	27	2.7
1Q4L	5	23	23.2	7	32	6.4	10	45	4.6
1Q5K	1	5	4.6	2	9	1.8	3	14	1.4
1Q41	3	14	13.9	7	32	6.4	7	32	3.2
1R0E	4	18	18.6	8	36	7.3	11	50	5.0
1UV5	3	14	13.9	7	32	6.4	7	32	3.2
2JLD	1	5	4.6	7	32	6.4	10	45	4.6
2O5K	4	18	18.6	5	23	4.6	7	32	3.2
2OW3	7	32	32.5	11	50	10.0	13	59	5.9
3DU8	1	5	4.6	6	27	5.5	9	41	4.1
3F7Z	3	14	13.9	7	32	6.4	10	45	4.6
3F88	2	9	9.3	6	27	5.5	8	36	3.6
Ensemble	6	27	27.9	13	59	11.8	16	73	7.3

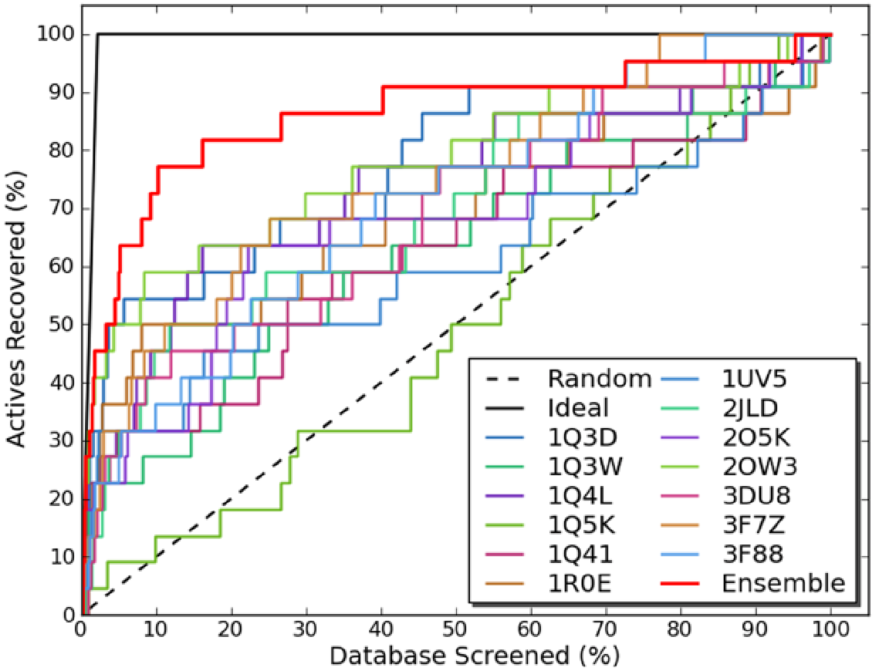
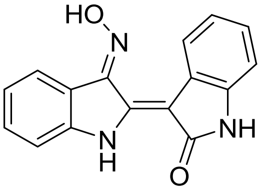
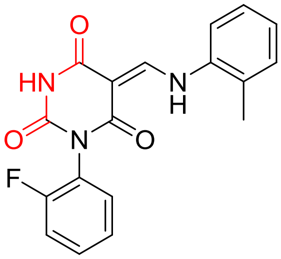
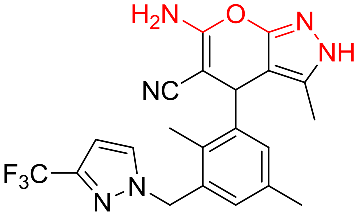
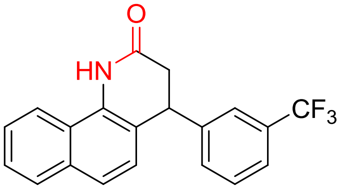
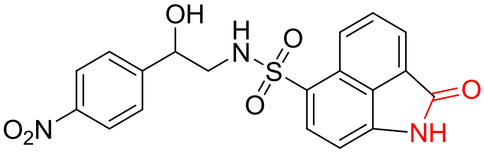
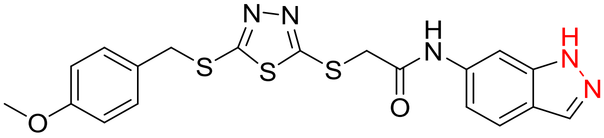


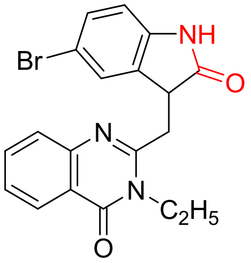
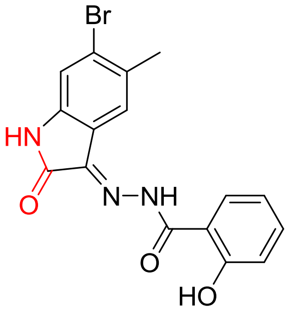
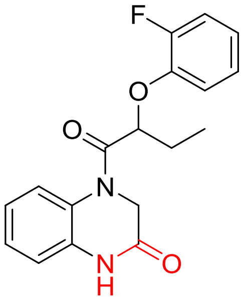
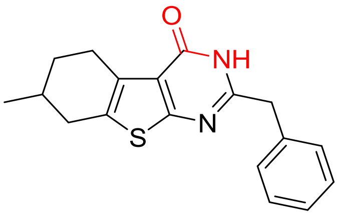
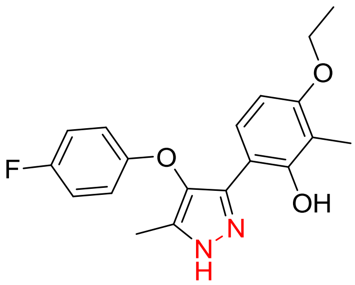
Figure 3.6. Enrichment Curve for the individual docking into different PDB structures (PDB labels given) and for ensemble docking.

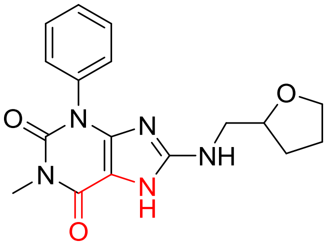
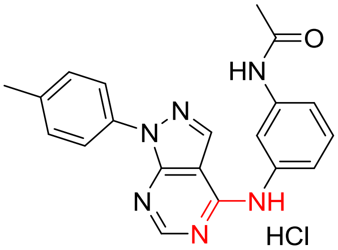
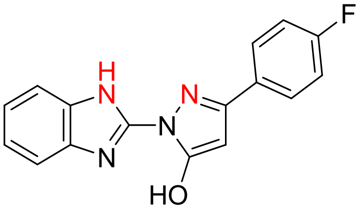
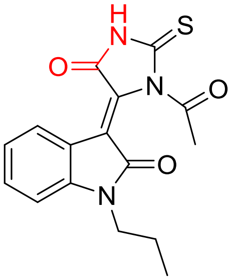
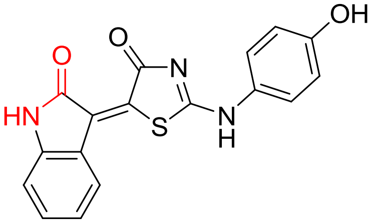
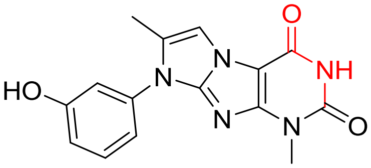
3.3.3. VIRTUAL SCREENING AND BIOLOGICAL VALIDATION

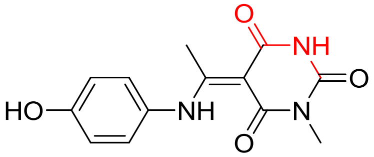
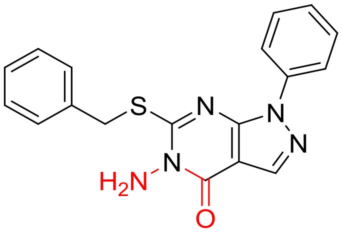
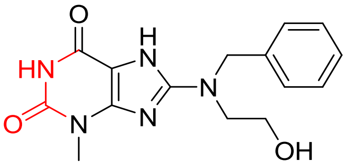
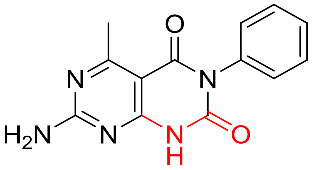
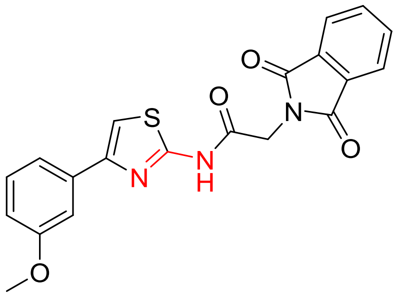
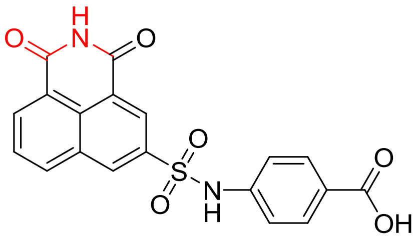
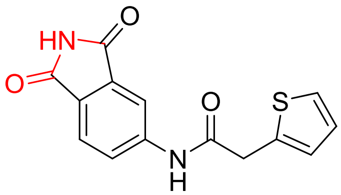
To verify the reliability of the proposed mixed ligand/structure-based approach in the virtual screening procedure from the experimental perspective, 24 hit compounds were selected based on the ensemble docking score and on visual inspection, in which we verified the presence in the docking poses of the essential hydrogen bond interactions with the hinge region of the receptor. The compounds were purchased from ChemBridge Corp. and the GSK-3β inhibitory activities of these compounds were tested according to the experiments described above. Their chemical structures and biological activities are listed in Table 3.9. The functional groups highlighted in red are predicted to be the moieties that establish hydrogen bond interactions with the residues in the hinge region of the receptor. Fifteen structurally diverse compounds with inhibition greater than 50% at 10 μM concentration were selected for IC_{50} determination. Among them, 9 compounds had IC_{50} values less than 10 μM. Furthermore, compound **23**, with molecular weight of 286 and logP of 1.68, exhibited sub-micromolar activity, and hence can be used as a good starting point for further drug development.

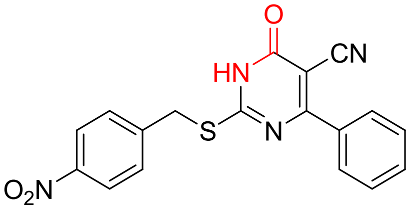
Table 3.9. Biological activities for binding to GSK-3β of the hit compounds.

No.	Structure	Docking Score (PDB code)	%inhibition at 10 μM (%)	IC_{50} (μM)
reference			100	0.19
1		–9.119 (2OW3)	13	N.A.[a]
2		–9.121 (1Q3D)	44	N.A.[a]
3		–9.433 (1UV5)	12	N.A.[a]
4		–9.401 (1R0E)	30	N.A.[a]
5		–9.149 (1Q5K)	65	39.51

6		–8.818 (1Q4L)	25	N.A.[a]
7		–8.923 (1Q4L)	100	15.79
8		–8.941 (1UV5)	34	N.A.[a]
9		–9.241 (1UV5)	25	N.A.[a]
10		–9.297 (1Q4L)	41	N.A.[a]

11		–10.566 (1Q5K)	84	2.06
12		–9.381 (1Q41)	65	67.07
13		–10.269 (1Q41)	30	N.A.[a]
14		–10.682 (1UV5)	75	1.05
15		–10.236 (1UV5)	100	1.76
16		–10.021 (1R0E)	100	2.40

17		–9.511 (1UV5)	61	72.73
18		–9.876 (1Q4L)	85	7.48
19		–9.716 (1R0E)	83	27.32
20		–9.752 (1Q41)	69	5.13
21		–9.571 (1Q4L)	75	10.02
22		–10.067 (2JLD)	100	2.80
23		–9.461 (1Q41)	100	0.51

24		–8.839 (2OW3)	65	1.58

[a] N.A.: not active.

To examine the predicted ligand-protein interactions, the best docking poses of the four most active hit compounds (compounds **14**, **15**, **23**, and **24**) are demonstrated in Figure 3.7. All four of the compounds satisfy the basic requirements of ATP-competitive GSK-3β inhibitors, which include functional groups which interact with the residues in the hinge region and an aromatic group which fits into the narrow ATP binding pocket. Furthermore, they all have additional interactions either with co-crystallized water molecules or with other protein residues. For instance, **14**, which was docked into the structure 1UV5, has a hydrogen bond interaction with one co-crystallized water molecule located on the bottom of the ATP-binding pocket through the carbonyl group on the indolone moiety; **15**, which was docked into the same crystal structure, has two hydrogen bonding interactions: one is between the hydroxyl group and the residue Asn64 and the other one is between the sulfur on the thiazole ring and the co-crystallized water; the carbonyl group on the amide moiety of **23**, which was docked into the structure 1Q41, has a hydrogen bonding interaction with one co-crystallized water molecule; while the cyano group of **24**, which was docked into the structure 2OW3, has a hydrogen bonding interaction with a different water molecule deeply buried in the ATP-binding pocket. These additional hydrogen bond interactions explained the high activities of the corresponding hit compounds.

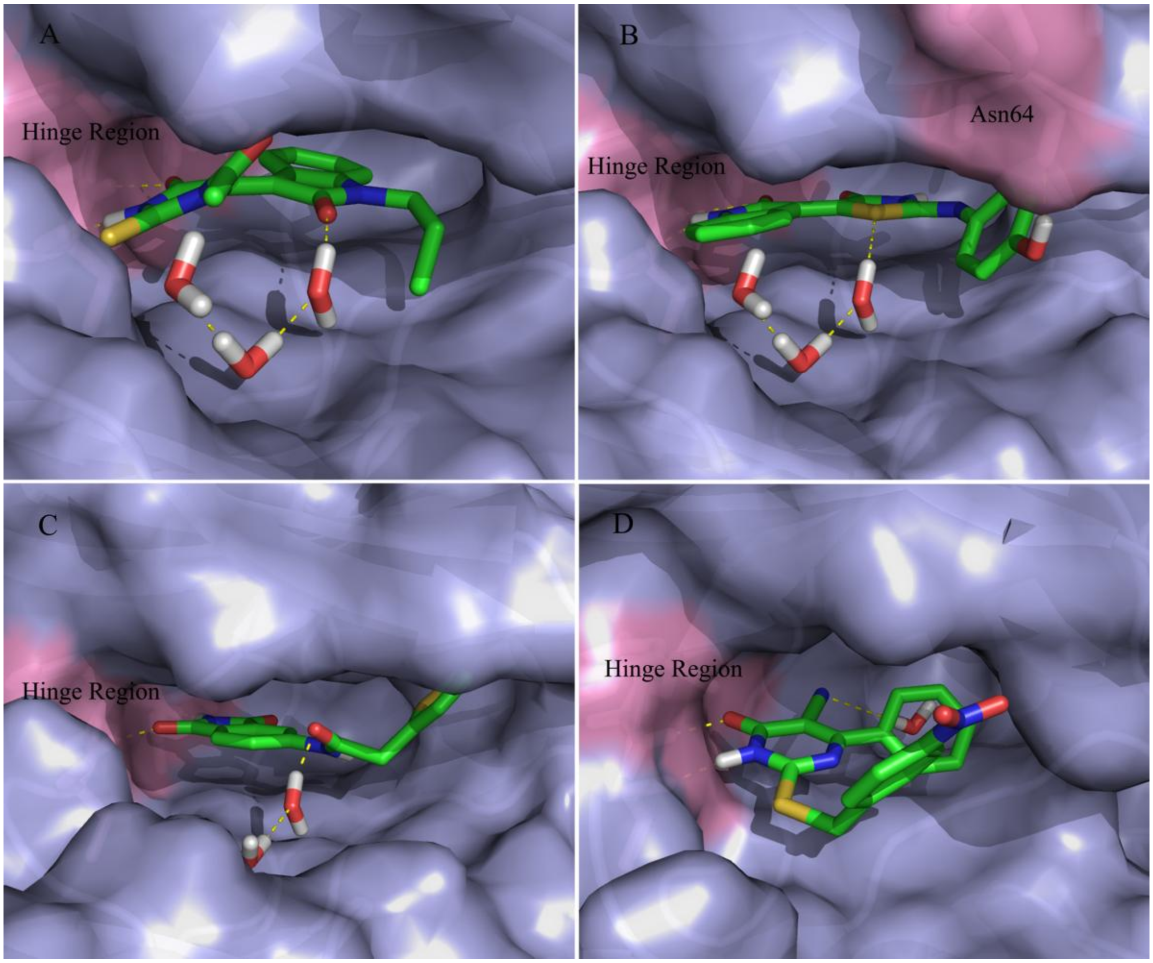


Figure 3.7. Docking poses of the four top active hit compounds. A. **14**; B. **15**; C. **23**; D. **24**. The protein surface is colored light blue, while the surface of the residues which have hydrogen bond interactions with the hit compounds are colored pink. The hit compounds and the co-crystallized waters have the same color codes: green (carbon), red (oxygen), blue (nitrogen), yellow (sulfur), and white (hydrogen).

3.4. CONCLUSIONS

In order to identify structurally novel and diverse GSK-3β inhibitors, we designed and validated a mixed ligand/structure-based virtual screening protocol. In terms of ligand-based approaches, we constructed a common pharmacophore hypothesis to be used for preliminary screening of large databases to make sure to include only compounds containing the key structural features needed to be GSK-3β inhibitors, and we performed diversity analysis to achieve maximum coverage of the activity space and reduce the structural redundancy in the screening process. In terms of a target-based approach, we systematically investigated various docking protocols and showed that ensemble docking is an efficient and effective technique to address the induced-fit effects in ligand-protein recognition. The hit compounds obtained from virtual screening underwent experimental validation. The bioassay results showed that 15 out of 24 hit compounds are indeed GSK-3β inhibitors when tested at 10 μM, and among them, 8 molecules exhibited low micromolar activities ($IC_{50} < 10$ μM) and one molecule exhibited sub-micromolar activity. The high hit rate demonstrated the success of the proposed mixed ligand/structure-based virtual screening protocol. Upon closely examining the predicted binding poses, we concluded that the additional hydrogen bond interactions with co-crystallized water molecules in the ATP binding pocket are required for the high binding affinity.

Chapter 4. HIERARCHICAL QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP ANALYSIS AND VIRTUAL SCREENING STUDIES FOR GLYCOGEN SYNTHASE KINASE-3

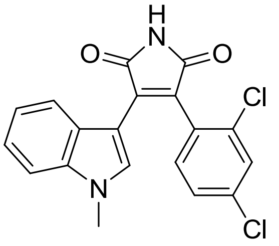
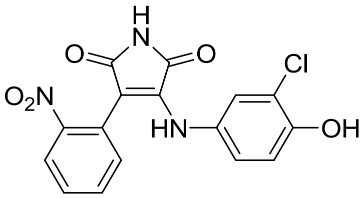
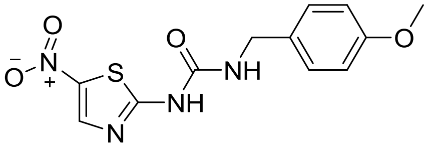
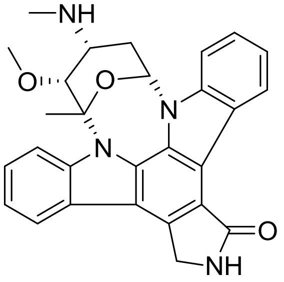
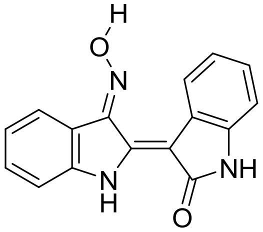
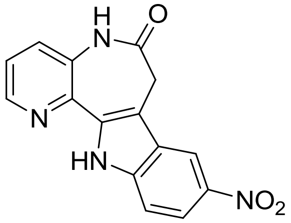
Gang Fu, Olivia R. Dale, Sheng Liu, Xiaofei Nan, Zhendong Zhao, Haining Liu, Yixin Chen, Dawn Wilkins, Susan P. Manly, Stephen J. Cutler, and Robert J. Doerksen

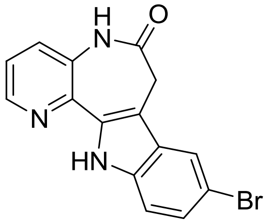
4.1. INTRODUCTION

Since 3D-QSAR using CoMFA and CoMSIA methods requires structural alignment, which is crucial and requires the molecules to have a similar scaffold. So the size of data set in one particular modeling project is typically limited. And sometimes, the conformations used for alignments are different from the bioactive conformations, which can reduce the accuracy and relevance of the model. In contrast to 3D-QSAR methods, classical QSAR methods based on structural and physicochemical descriptors are independent of structural alignment, so they can be expected to perform well with large data sets. Furthermore, with the development of machine learning algorithms and artificial intelligence methods which can be implemented for both model construction and feature selection, modern QSAR methods which can produce highly predictive linear or nonlinear models play an increasingly important role in the drug discovery process, especially in virtual screening studies to identify novel hit compounds.[79]

The predictive performance of the QSAR models are highly dependent on the consistency of the experimental bioactivities, in other words, the data homogeneity. The different bioassay protocols can conceivably yield distinct biological activities for the same compound inhibiting GSK-3 (Table 4.1). The systematic errors from the experiments can significantly reduce the predictive accuracies of the QSAR models if compounds with biological activities determined under different bioassay protocols are collected into one large data set.[136] To deal with this circumstance, we constructed a hierarchical QSAR model which adopts a multi-level structure to take into account the data heterogeneity. Figure 4.1 illustrates the multi-level structure of the hierarchical QSAR model. In the lower level of the model, each regression model is built on data collected by a single research group using a particular bioassay. The labels to be used in building regression models are the single-protocol bioactivities. In the upper level of the model, all the compounds culled from different research groups are collected in a single data set, and a multi-class classification model is constructed to separate compounds into different subclasses. The labels of the classification models are the group numbers (from I to VII) which are used to indicate the heterogeneity in the data. State-of-the-art machine learning algorithms, support vector machines (SVM) and random forests (RF), were used to construct a multi-class classification model at the higher level and multiple regression models at the lower level. The performance of the two algorithms was systematically investigated and compared. The best models with the highest predictive ability were employed, combined with ensemble docking, in a mixed ligand-based/structure-based virtual screening study to identify structurally novel GSK-3 inhibitors.

Table 4.1. GSK-3 inhibitors and their different reported biological activities.

Structure	Biological Activities
 SB-216763	Coghlan et al.[137] IC_{50} = 34.3 nM Kozikowski et al.[138] IC_{50} = 50 nM
 SB-415286	Coghlan et al.[137] IC_{50} = 77.5 nM Smith et al.[61] IC_{50} = 104±10 nM Kozikowski et al.[138] IC_{50} = 1.3 μM
 AR-A014418	Bhat et al.[42] IC_{50} = 104±27 nM Gaisina et al.[73] IC_{50} = 41.8±6.4 nM
 Staurosporine	Leclerc et al.[62] IC_{50} = 15 nM Engler et al.[139] IC_{50} = 56±6.9 nM
 Indirubin-3’-monoxime	Meijer et al.[44] IC_{50} = 22 nM Bain et al.[140] IC_{50} = 0.19 μM
 Alsterpaullone	Leost et al.[141] IC_{50} = 4 nM Bain et al.[140] IC_{50} = 0.11 μM

 Kenpaullone	Leost et al.[141] IC_{50} = 23 nM Bain et al.[140] IC_{50} = 0.23 μM

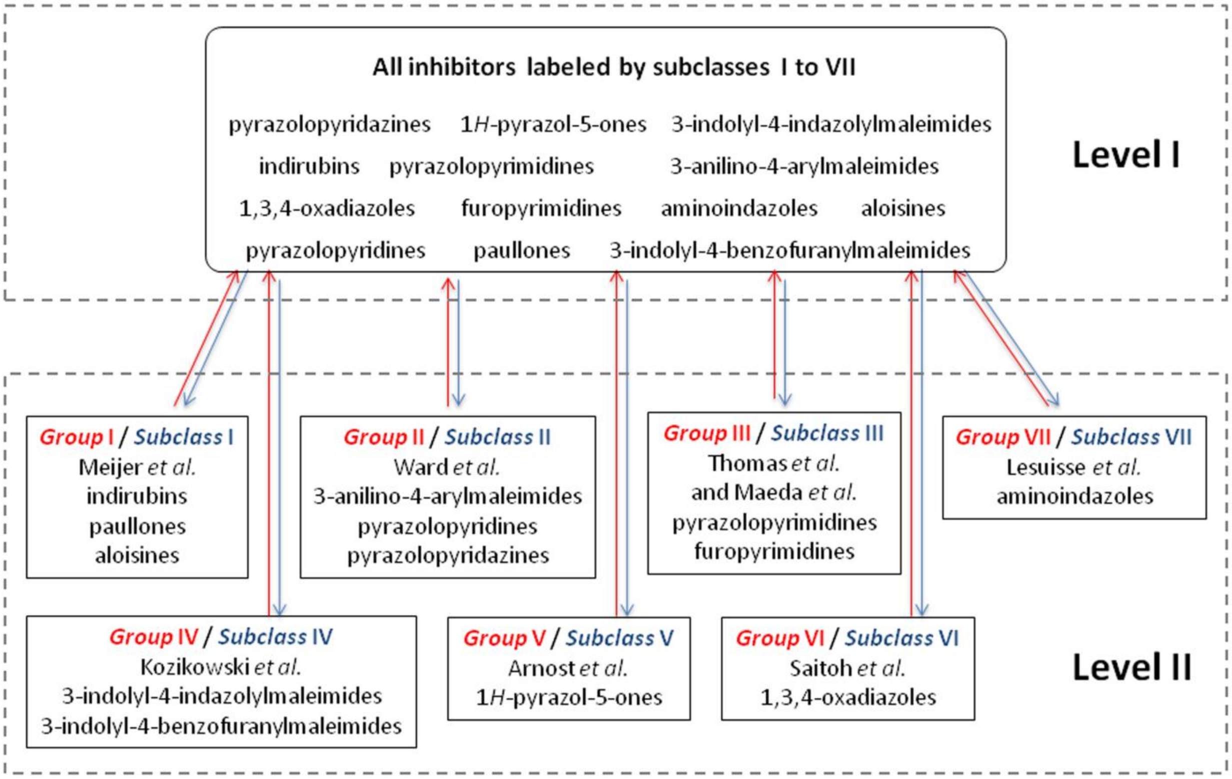


Figure 4.1. The multi-level structure of the hierarchical QSAR model. Level I constitutes a multi-class classification model and Level II constitutes regression models. The group numbers highlighted in red were used in model construction, indicating the data heterogeneity. The subclass numbers highlighted in blue were used in the classification models, indicating the corresponding groups.

4.2. MATERIALS AND METHODS

4.2.1. COLLECTION OF HETEROGENEOUS DATA SETS

A collection of seven groups of compounds were compiled from the literature. To demonstrate the difference of the bioassay protocols under which the inhibitory activities of each group of compounds was tested, we describe and compare some experimental details here, especially regarding the isoforms and organisms of the target proteins, the substrate peptide, and the concentration of ATP cofactors. For each dataset we included from the corresponding papers all the reported compounds except for those without reported activity and or without well-defined stereochemistry. The chemical structures of the collected 728 GSK-3β inhibitors from seven research groups, associated with the experimental and predicted pIC_{50} values, are provided (Appendix: B-H).

Meijer *et al.* studied the structure-activity relationships (SAR) of hundreds of compounds against GSK-3α/β. The compounds typically belonged to three chemical classes, indirubins,[62, 63, 142] paullones,[64, 143] and aloisines.[65] A total of 214 of the molecules constitute the Group I dataset. The GSK-3α/β was purified from porcine brain by affinity chromatography on immobilized axin and purified from insect Sf9 cells. The kinase activity was assayed using the pGS-1 peptide as a substrate, at a final ATP concentration of 15 μM.

Ward *et al.* studied the binding affinity of several sets of compounds again human GSK-3α (hGSK-3α). The compounds defined as the Group II dataset consist of the analogs of 3-anilino-4-arylmaleimides,[61] the analogs of 5-aryl- pyrazolopyridines and 5-aryl-pyrazolopyridazines,[68, 72] and the analogs of 6-aryl- and 6-heteroaryl-pyrazolopyridines.[70, 71] A total of 157 molecules were collected. The hGSK-3α isoform was expressed using the baculovirus expression system and recovered from cells by homogenization and purification using NiNTA superflow. The hGSK-3α was assayed on the substrate peptide (Biotin-KYRRAAVPPSPSLSRHSSPHQSpEDEEE) in the presence of 10 μM ATP.

Thomas *et al.* reported the *in vitro* biological activities for the inhibition of human GSK-3β (hGSK-3β) by the pyrazolopyrimidine derivatives,[66, 67] and Maeda *et al.* reported the same inhibitory activities by the furopyrimidine derivatives.[112] A total of 91 molecules based on these two chemotypes constitute the Group III dataset. The hGSK-3β enzyme assays for the two sets of compounds were carried out by Schweiker *et al.* from GlaxoSmithKline Inc. under the same scintillation proximity assay (SPA) protocol. The hGSK-3β was assayed on the substrate peptide (Biotin-Ahx-AAAKRREILSRRPSpYR-amide) in the presence of 2.5 μM ATP.

Kozikowski *et al.* investigated the SAR of a set of potent and selective GSK-3β inhibitors based on the chemotypes of 1*H*-indazol-3-yl-(indol-3-yl)maleimides and benzofuran-3-yl-(indol-3-yl)maleimides.[73, 138, 144] A total of 85 of the molecules gathered from the literature define the Group IV dataset. The *in vitro* kinase assay was performed on the pGS peptide (RRRPASVPPSPSLRHSSpHQRR) in the presence of 10 μM ATP.

Arnost *et al.* identified a series of 3-aryl-4-(arylhydrazono)-1*H*-pyrazol-5-ones as potent GSK-3β inhibitors,[52] of which a total of 62 molecular derivatives constitute the Group V dataset. The compounds were tested against GSK-3β using the standard coupled enzyme assay, which was carried out in the presence of substrate peptide (HSSPHQSpEDEEE) and 10 μM ATP.

Saitoh *et al.* reported the design, synthesis and SAR analysis of a novel series of 1,3,4-oxadiazole derivatives as GSK-3β inhibitors.[49, 50] A set of 61 of the compounds exhibiting good potency and selectivity constitute the Group VI dataset. The human GSK-3β expressed using baculovirus expressing system was assayed in the presence of 0.5 μM ATP.

Lesuisse *et al.* reported the rational design of potent and selective GSK-3β inhibitors via structural modifications to lead compounds identified through high throughput screening.[97] The lead compound contains an aminoindazole moiety interacting with the hinge region of the kinase binding pocket, and a total of 58 derivative molecules constitute the Group VII dataset. The enzymatic activity was measured by the SPA approach, in which the recombinant human GSK-3β was assayed on pGS-2 substrate peptide in the presence of 1 μM ATP.

4.2.2. CALCULATION OF MOLECULAR DESCRIPTORS

The 3D structures were built and preoptimized using the LigPrep module[132] from Schrödinger Suite 2010. In order to obtain reasonable conformers of the compounds which were to be employed to calculate relevant 3D descriptors, ensemble docking[145] was performed to generate the predicted binding poses for all the compounds in the QSAR study. The resulting conformations were found to be consistent for structurally similar compounds and are close to biologically significant conformers, comparing to the co-crystallized structures. The 3D structures were further optimized until the root-mean-square gradient reached 0.0001 kcal mol^{-1} and partial charges were obtained with the semiempirical AM1 method implemented in MOE software.[146] The optimized structures associated with the partial charges were then subjected to molecular descriptor calculations using DragonX software[147] which generates up to 3224 descriptors (Table 4.2). In order to remove descriptors that would likely be unhelpful for generating high quality models, several preprocessing steps were performed, including the elimination of descriptors with the same values for all molecules in the training set for the model being built, descriptors with too many zero values (>90%), and descriptors with very small standard deviations (0.05). After preprocessing, the feature values were scaled to have 0 as mean and 1 as standard deviation before they were applied in the SVM modeling. The normalization step is necessary for SVM, since the ranges of feature values vary greatly from one category to another, and the large variations have a big impact on the quality of SVM models. However, the normalization step is not required for RF, so the initial feature values were used in the RF modeling.

Table 4.2. Twenty-two Categories of molecular feature descriptors generated using DragonX software.

2D or 3D	Category	Number of descriptors
2D	constitutional descriptors	48
	topological descriptors	119
	walk and path counts	47
	connectivity indices	33
	information indices	47
	2D autocorrelations	96
	edge adjacency indices	107
	Burden eigenvalue descriptors	64
	topological charge indices	21
	eigenvalue-based indices	44
	functional group counts	154
	atom-centered fragments	120
	molecular properties	29
	2D binary fingerprints	780
	2D frequency fingerprints	780
3D	Randic molecular profiles	41
	geometrical descriptors	74
	RDF descriptors	150
	3D-MoRSE descriptors	160
	WHIM descriptors	99
	GETAWAY descriptors	197
	charge descriptors	14

Table 4.3. Data set summary.

Group	Research Group	Chemical Classes	N_{train}	pIC_{50} range (train)	N_{test}	pIC_{50} range (test)
I	Meijer *et al.*	Indirubins Paullones Aloisines	174	3.0-8.5	40	4.6-8.0
II	Ward *et al.*	3-Anilino-4-arylmaleimides 5-Aryl- pyrazolopyridines 5-Aryl-pyrozolopyridazines 6-Aryl-pyrazolopyridines 6-Heteroaryl-pyrazolopyridines	128	5.3-9.1	29	5.3-8.3
III	Thomas *et al.* Maeda *et al.*	Pyrazolopyrimidines Furopyrimidines	75	4.5-8.8	16	4.5-8.4
IV	Kozikowski *et al.*	1*H*-Indazol-3-yl-(indol-3-yl)maleimides Benzofuran-3-yl-(indol-3-yl)maleimides	66	4.9-9.6	19	5.6-9.3
V	Arnost *et al.*	3-Aryl-4-(arylhydrazono)-1*H*-pyrazol-5-ones	49	5.4-9.4	13	6.6-8.7
VI	Saitoh *et al.*	1,3,4-Oxadiazoles	49	5.0-8.6	12	6.1-8.2
VII	Lesuisse *et al.*	Aminoindazoles	46	4.0-8.3	12	4.6-8.0

4.2.3. DIVISION INTO TRAINING AND TEST SETS

K-mean clustering analysis was employed in the division of each group of compounds into the training and test sets (Table 4.3).[148] As an unsupervised learning algorithm, the method partitions the data set into *k* mutually exclusive clusters, with *k* any integer, and from each cluster, the compound with the median biological activity was selected for the test set and the rest of the compounds in the cluster were selected for the training set. Hence, the number of clusters defines the number of compounds in the test set, which is usually about one fifth of the overall data set for each group. Furthermore, such a split guarantees that the distribution of biological activities for the test set is similar to that for the training set. The feature values were scaled to the range [0, 1] before they were used to calculate the squared Euclidean distance, which then was minimized for each cluster during the k-mean analysis. For each group of compounds, the same division into training and test sets was applied for both the multi-class classification and regression models.

4.2.4. MEASUREMENT OF THE PREDICTION PERFORMANCE

For the multi-class classification models, the quality of the models was estimated using the classification accuracy.

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ compounds}{Total\ number\ of\ compounds} \times 100\% \qquad (4.1)$$

Confusion tables were also used to illustrate the prediction ability of the models in more detail.

For the regression models, three statistical measures were employed to evaluate the predictive accuracies of the models, including the squared Pearson’s correlation coefficient (R^2), residual mean square error (RMSE), and F-ratio. The goodness-of-fit R^2 is defined by:

$$R^2 = \frac{(n\sum \hat{y}_i y_i - \sum \hat{y}_i \sum y_i)^2}{(n\sum \hat{y}_i^2 - (\sum \hat{y}_i)^2)(n\sum y_i^2 - (\sum y_i)^2)} \qquad (4.2)$$

where y_i represents the observed bioactivities and $\hat{y}_i$ represents the predicted bioactivities. Meanwhile, the subscript representations are adopted to indicate whether the relevant statistic refers to the training set (R^2_{train}), leave-one-out cross-validation (R^2_{LOO}), 5-fold cross-validation (R^2_{CV}), or prediction of the test set (R^2_{test}).

The RMSE or residual variance is calculated as follows:

$$RMSE = \frac{\sum(\hat{y}_i - y_i)^2}{n-2} \tag{4.3}$$

where $RMSE$ is equivalent to s_{res}^2, the squared standard deviation of the estimate. The same subscript representations are adopted as above. The squared deviation of the regression line (s_r^2) is defined as:

$$s_r^2 = \sum(y_i - \bar{y})^2 \tag{4.4}$$

where $\bar{y}$ represents the average of the observed bioactivities. The s_{res}^2 and s_r^2 combined together can be used to calculate the F-ratio, according to the equation:

$$F = \frac{s_r^2}{s_{res}^2} \tag{4.5}$$

where F has the degree of freedom 1 and n–2 for s_r^2 and s_{res}^2, respectively. This ratio is used in hypothesis testing with the assumption that the model predicts better than a prediction using an average value, α, at a certain significance level (α = 0.05). So the higher the value of the F-ratio, the better the predictiveness of the model.[149] The same subscript representations are adopted as above.

The robustness and predictive power of the regression models were further examined using the Y-randomization test which is a widely used validation technique. The test is carried out by rebuilding the new regression models using the original independent variable matrix (the feature values) but using a randomly permuted dependent variable vector (for the bioactivities). This process is repeated 100 times, and the mean values and standard deviations of the measurements of the new models are evaluated and compared with the original models. It is expected that the new models obtained with the randomized bioactivities should generally yield poor prediction performance in terms of both internal and external validation. If the best resulting models in the Y-randomization test exhibit high prediction ability, then it implies that a reliable regression model cannot be obtained for the given data set using the current modeling method. The Y-randomization test was applied to the seven regression models using the selected methods.

4.2.5. SUPPORT VECTOR MACHINES

SVM have been developed by Vapnik[150] and his co-workers, based on the statistical learning theory derived from the structural risk minimization principle and Vapnik-Chervonenkis (VC) dimension. SVM involves the attempt to minimize the upper bound of the expected test error, and it is superior in terms of generalization over the traditional empirical risk minimization principle employed in conventional neural networks, which just minimize the error on the training data. SVM was initially developed in the context of binary (two-class) classification problems, in which a linear classifier with the smallest empirical risk and VC dimension is constructed to correctly separate the data into either the positive or negative class. Geometrically, this classifier is termed the optimal separating hyperplane with the maximum margin for a given set of learning data. The largest margin corresponds to the smallest empirical risk and smallest VC dimension.

The current standard formulation is soft margin SVM[151] which introduces a penalty for misclassified data, denoted by ξ and called a slack variable. In the case of binary classification, given *n* points of training data placed in the matrix $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})$, where each point $\mathbf{x_i}$ is a *m*-dimensional real vector, $\mathbf{x_i} \in R^m$, for $i = 1, 2, \ldots, n$, and each object $\mathbf{x_i}$ belongs to a class $y_i \in \{-1, +1\}$, the construction of a linear classifier actually is equivalent to solving the following primal optimization problem:

$$\min \frac{\|\boldsymbol{\omega}\|^2}{2} + C \sum_{i=1}^{m} \xi_i \tag{4.6}$$

$$\text{with the constraints } \begin{cases} y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) \geq +1 - \xi_i, & \forall i \\ \xi_i > 0, & \forall i \end{cases} \tag{4.7}$$

where $\boldsymbol{\omega}$ is a norm vector perpendicular to the hyperplane, and C is a tuning parameter that can be adjusted by the user. Minimizing the first term can be translated into finding the maximum margin hyperplane and minimizing the second term can be translated into minimizing the error of misclassified data. So the tuning parameter C represents a trade-off between the expected test error, or generalization error, and the error of the misclassified learning data, or training error. This is a quadratic programming problem, which minimizes a quadratic function under linear constraints. The problem can be solved based on the use of a Lagrange function, transforming the primal problem into its dual formulation. After solving the dual problem, the optimal values for the vector $\boldsymbol{\omega}$ and threshold b can be used to construct an optimal separating hyperplane for the classification of new data.

In the case of multi-class classification, the ‘one-versus-one’ approach is involved. This approach first decomposes the training set into several binary classification problems, and then trains a binary classifier for every two-class problem from the training set. So a *k*-class problem would result in *k**(*k*–1)/2 binary classifiers. In the prediction phase, a voting strategy is engaged to assign the class of the data points and a data point is designated to be in the class with the maximum number of votes. If two classes have an identical number of votes, the data point is assigned to the class appearing first in the storage array.

In the case of hard-to-separate classes, a nonlinear kernel function $\phi(\mathbf{x_i})$ is introduced in order to map the original feature space into a feature space of a higher dimension, and then a linear classification is performed in that higher dimensional feature space. By transforming the feature space, the nonlinear separable classes having a complex relationship between the original input variables ($\mathbf{x_i}$) and the corresponding classes (y_i) can be efficiently separated by a linear classifier. One of the most widely used kernel function, the radial basis function (RBF) kernel in the Gaussian form, was used in the present study:

$$K(x_i, x_j) = exp\left(-\gamma\left\|x_i - x_j\right\|^2\right), \gamma > 0 \tag{4.8}$$

where the tuning parameter γ controls the shape of the separating hyperplane. It can be optimized with a suitable cross-validation procedure.

Initially developed for classification, SVM was extended by Vapnik and his co-workers[152] to be able to establish regression models. The transformation from a sign function in SVM classification to a real function in SVM regression is realized by incorporating an ε-insensitive loss function:

$$L_\epsilon(y_i, f(\mathbf{x_i})) = \begin{cases} 0, & if\ |y_i - f(\mathbf{x_i})| \leq \varepsilon \\ |y_i - f(\mathbf{x_i})| - \varepsilon, & otherwise \end{cases} \tag{4.9}$$

where ε represents the maximum margin for the deviation between the experimental values (y_i) and predicted values ($f(\mathbf{x_i})$), and it should be optimized during the model construction. Based on this linear loss function, the loss penalty for the deviation within ε is zero, and the loss penalty for the deviation beyond ε is calculated by the difference between the deviation and ε. The loss penalty corresponds to the slack variable in the soft margin SVM classification models. However, in regression models, overestimation and underestimation cause distinct prediction errors. So two slack variables are introduced, which are ξ^+ and ξ^-, and the primal formulation for soft margin SVM regression models is:

$$\min \frac{\|\omega\|^2}{2} + C\sum_{i=1}^{m}(\xi_i^+ + \xi_i^-) \tag{4.10}$$

$$\text{with the constraints} \begin{cases} f(\mathbf{x_i}) - y_i \geq \varepsilon + \xi_i^+, & \forall i \\ y_i - f(\mathbf{x_i}) \leq \varepsilon + \xi_i^-, & \forall i \\ \xi_i^+, \xi_i^- > 0, & \forall i \end{cases} \tag{4.11}$$

In the present work, the Matlab[153] interface of LIBSVM 3.0[154] was employed for the SVM classification and regression analysis using a Gaussian RBF kernel in the hierarchical QSAR modeling. There are two tuning parameters for SVM classification models (C, γ) and three tuning parameters for SVM regression models (C, γ, ε), which balance the trade-off between data fit and model complexity. The tuning parameters were determined by a grid search. 100 replications of 5-fold cross-validation were performed to access the classification accuracies or regression R_{CV}^2 at each point over a fixed grid of parameter values. The median values for the 100 replications were used as the optimal tuning parameters.

In order to demonstrate the influence of redundant or irrelevant features to SVM, the computationally intensive wrapper-based feature selection method was implemented. Wrapper methods incorporate model assessment within the feature selection procedure, which usually provides superior performance and can be used to find the optimal or suboptimal subset of features specifically for certain learning models,[155] such as for multiple linear regression, artificial neural networks, or SVM. Since the number of subsets of features is 2^M where M is the total number of features, there is no way to do an exhaustive search. However, randomized heuristic search methods provide a promising feature selection mechanism, especially when implemented with evolutionary algorithms from artificial intelligence, such as the particle swarm algorithm used in the present study. After the wrapper-based feature selection, one post-processing step was employed in order to further reduce the redundancy and refine the models, which was to eliminate features which are highly correlated (Pearson's correlation coefficient R > 0.9) with others and hence would be expected to be of low importance to the models.

4.2.6. BINARY PARTICLE SWARM ALGORITHM

The particle swarm (PS) algorithm is a part of evolutionary computing, which belongs to artificial intelligence. The algorithm, introduced by Kennedy and Eberhart in 1995,[156, 157] simulates the collective behavior of a flock of birds or a school of fish looking for the shortest route from their current position to a food source. PS optimization, like other swarm intelligence algorithms such as ant colony optimization, is a stochastic, population-based optimization algorithm which explores the search space without the provision of a global model. The randomly initialized particles with an original velocity and position proceed through the search space, remembering the best position encountered; then they accelerate towards the position of the best performing particles as well as towards their personal best previous position. So it is a valuable heuristic algorithm updating the velocities and positions according to the historic behaviors of the particles themselves or of their neighbors. During each update, the velocity is adjusted by two elements, the cognitive part and social part. Later, Shi and Eberhart proposed another improved model, introducing the inertia weight to achieve a fast convergence.[158, 159] The position is adjusted using the updated velocity.

The binary particle swarm representation[160] was used in the present study, which is widely used to search the feature space in modern QSAR studies.[161, 162] The features are encoded in a population (P) of binary strings, indicating the positions of particles. Each binary string represents a point in M-dimensional space, in which M is the total number of candidate features. The position of the ith particle is represented by a vector X_i = (x_{i1}, x_{i2}, …, x_{iM}), where x_{id} (i = 1, 2, …, P and d = 1, 2, …, M) can be either 0 representing an unselected feature or 1 representing a selected feature. The velocity of the ith particle is represented by a number v_i, where v_i (i = 1, 2, …, P) can be a positive integer, varying between 1 and $V_{\max}$, based on how many bits in the string should be changed from 0 to 1 or from 1 to 0 in each iteration. The maximum velocity $V_{\max}$ serves as a constraint to control the trade-off between the local and global exploration of a particle swarm. According to Wang *et al*,[160] the $V_{\max}$ should be neither too low nor too high, and a moderate value such as M/3 is recommended. The velocity can be initialized by a random number generated from a uniform distribution from 1 to M, and if the velocity is larger than $V_{\max}$, it will be set back to the maximum velocity. The positions of the particles are initialized by probabilistic selection. A set of random numbers between 0 and 1 are first generated from the uniform distribution. Then probability thresholds (δ) are used to convert continuous numbers to a binary representation, via mapping the random numbers in the interval (0, δ) to 1 and other values to 0. Since there are thousands of candidate features, and only a few of them (10-40) are expected to be selected in a particular QSAR model, a small threshold interval (<0.1) is used to determine the subset membership.

Once the particles are constructed, the velocity and position of each particle are updated on every iteration according to the following equations:

$$v_i = \omega \times \ v_i \ + \ c_1 \times \mathrm{rand}_1 \times (p_{id} - x_{id}) + c_2 \times \mathrm{rand}_2 \times \left(p_{gd} - x_{id}\right) \tag{4.12}$$

$$x_{id} = x_{id} + v_{id} \tag{4.13}$$

where c_1 and c_2 are cognitive and social acceleration constants, contributing to the balance between local and global exploration; rand_1 and rand_2 are two random numbers, generated in the range (0, 1) from a uniform distribution; p_{id} represents the best previous position of the ith particle, and p_{gd} represents the overall best position of all the particles in the population. The inertia weight (ω) is introduced to limit the number of iterations needed for convergence. It is linearly decreased along with the iteration (k) for each update:

$$\omega_{k+1} = \omega_{max} - \frac{\omega_{max} - \omega_{min}}{k_{max}} \times k \tag{4.14}$$

where ω_{max} and ω_{min} are predefined maximum and minimum values of inertia weight, and k_{max} is the predefined maximum number of iterations. The starting point ω_1 equals ω_{max}. After the predefined number of iterations, the inertia weight reduces to ω_{min}.

Within each iteration, a fitness value indicating the goodness of the particular subset of features is computed for every particle as follows:

$$Fitness = \begin{cases} \mu_{CV}, & if\ m < \alpha \\ \mu_{CV}{}^{\frac{m}{\alpha}}, & if\ m \geq \alpha \end{cases} \tag{4.15}$$

where μ_{CV} is the measure of goodness describing how well the model fits the observation. For multi-class SVM classification models, it is calculated based on the median classification accuracy for 9 replications of 5-fold cross validation; and for SVM regression models, it is calculated based on the median R_{CV}^2 for the 9 replications of 5-fold cross validation. The actual number of selected features is represented by m and the desired number of selected features is represented by α. Since a smaller number of selected features give higher interpretability and generalization power to the model, the optimal particle position is the feature subset with the shortest length and the highest measure of goodness of fit. The piecewise fitness function employs the exponential function to penalize solutions containing a large number of selected features.

After preliminary studies (results not shown), the tuning parameters controlling the performance of the PS feature selection algorithm were determined (Table 4.4), including the population size (P), the maximum number of iterations (k_{max}), the cognitive acceleration constant (c_1), the social acceleration constant (c_2), the maximum value of inertia weight (ω_{max}), the minimum value of inertia weight (ω_{min}), the maximum velocity ($V_{\max}$), and the probability threshold (δ). Some of them were determined based on the recommended values according to Wang *et al*.[160] The tuning parameters for SVM were also defined for the feature selection process.

Table 4.4. The parameters controlling the performance of PS feature selection algorithm as well as SVM classification and SVM regression models.

Group	P	$k_{\max}$	c_1	c_2	$\omega_{\max}$	$\omega_{\min}$	$V_{\max}$	δ	C	γ	ε
					multi-class SVM classification						
–	500	2000	2.0	2.0	1.4	1.4	452	0.02	30	0.08	–
					SVM regression						
I	400	6000	2.0	2.0	1.4	0.4	428	0.05	50	0.05	0.08
II	500	2000	2.0	2.0	1.4	0.4	430	0.04	50	0.05	0.08
III	500	2000	2.0	2.0	1.4	0.4	383	0.03	30	0.05	0.06
IV	400	4000	2.0	2.0	1.4	0.4	406	0.04	30	0.05	0.06
V	400	3000	2.0	2.0	1.4	0.4	369	0.03	30	0.08	0.08
VI	400	3000	2.0	2.0	1.4	0.4	393	0.03	30	0.08	0.08
VII	400	3000	2.0	2.0	1.4	0.4	387	0.03	30	0.08	0.08

4.2.7. RANDOM FORESTS

Decision trees are commonly used for data mining. The most popular kind is classification and regression trees (CART), which is a greedy method based on a recursive partitioning algorithm. CART has a combination of several advantages over other machine learning algorithms including the ability to ignore irrelevant descriptors, the ability to handle mixed data structures (continuous and discrete variables), and the simplicity of the results which can be represented by simple tree models rather than by elaborate equations. CART can be used to construct nonparametric and nonlinear models in order to predict either continuous or categorical dependent variables. However, the major drawback of CART is its low prediction accuracy caused by the overfitted tree-based structure, especially when it is used to deal with a large dataset. To avoid overfitting, different approaches have been applied including computation-intensive pruning methods based on cross-validation or ensemble learning methods such as random forests (RF). Developed by Leo Breiman[163], RF has been demonstrated as one of the most powerful tools for data exploration, delivering improved prediction accuracy while retaining the appealing properties of decision tree methods.[164] The RF algorithm used in the present study is available in the Matlab implementation.[165]

The random forests are a collection of CART-like trees, which are grown from bootstrap samples of the original training data. A total number of n_{tree} trees are grown following the CART algorithm until the maximum size of each tree is reached, and the trees are not pruned back. The prediction of the new data is made based on the aggregated outputs of the ensemble trees. For classification problems it is the class with the majority of votes and for regression problems it is the value of the average prediction.

In the process of training, each CART-like tree is grown using a bootstrap sample of n molecules drawn from the training data with replacement. Since bootstrapping is random sampling with replacement, some of the molecules from the training data will appear multiple times in the bootstrap sample, while some others will be left out of the sample. The ‘left out’ molecules account for one-third of the training set molecules on average and they constitute the out-of-bag (OOB) sample. The OOB molecules which have not been used in tree construction will be used as the internal test set (validation set) to estimate the prediction performance. This internal OOB error estimate has been proven to be unbiased and in good agreement with k-fold cross-validation.[164] So there is no need to perform additional cross-validation which would be computationally intensive, to get an unbiased estimate of the RF performance.

Three tuning parameters can be optimized by users to boost the performance of RF models: the number of trees (n_{tree}), the minimum node size and the number of randomly selected subsets of descriptors used for splitting at each node (m_{try}). n_{tree} should be sufficiently large so that the OOB error estimate can be stabilized, and the default value of 500 is usually large enough. The minimum node size determines the minimum size of nodes below which no split will be applied, and it controls the maximum size possibly reached for each tree. The default value for classification is 1 and the default value for regression is 5. m_{try} ranges from 1 to M, the total number of features available, and it serves as a trade-off between the impacts of two factors: the correlation between any pair of trees in the forest and the strength of each individual tree in the forest. The default value of m_{try} for classification is $\sqrt{\mathrm{M}}$ and the default value for regression is $\mathrm{M}/3$. It has been found that the performance of RF models is insensitive to changes in the three tuning parameters, and the default values are good enough in most cases.[164] So the default values for the classification and regression model tuning parameters were used in the present study.

The RF algorithm can handle thousands of features via an embedded feature selection algorithm built into the process of model construction. The approach can serve to evaluate how much a single feature contributes to the prediction accuracy based on how many times the same feature is used in the ensemble of trees. With the intrinsic feature selection algorithm, RF is generally insensitive to the presence of redundant or irrelevant descriptors, and there is no need to perform extra feature selection algorithms such as those which should be implemented in SVM models. As an illustration, the commonly used feature reduction method recursive backward elimination (RBE) was used and the results were compared with the original model.

The method, introduced by Svetnik *et al.*,[164] is based on the ranking of feature importance calculated by the RF algorithm itself during model construction. In each step of reduction, the least important half of the features is removed. The 100 replications of 5-fold cross-validation on the training set were calculated to access classification accuracies or regression R_{CV}^2, and the best median values indicate the optimal number of features.

4.2.8. MIXED LIGAND/STRUCTURE-BASED VIRTUAL SCREENING

The commonly used QSAR-based virtual screening approaches usually employ QSAR models based on a single group of data, which limits the exploration of ligand-based structural information. With the employment of the hierarchical QSAR model described above, we can fully exploit the usefulness of the structural information of many groups of small molecules obtained through extensive literature search. The hierarchical QSAR model was built in a bottom-up way, from the lower level to the upper level, and it is employed in the virtual screening study in a top-down way. The compounds from the chemical database first undergo multi-class classification at the upper level. Once a compound is classified into a certain subclass, its bioactivity is predicted using one of the regression models at the lower level. The multi-level structure of the hierarchical QSAR model guarantees that the molecular bioactivities are predicted using the most reliable model.

In order to obtain reliable predictions, a QSAR model should be used within its applicability domain,[166] which defines the acceptable interpolation regions in the multivariate space. The interpolation estimates the values within the endpoints. In one dimensional space, an interpolation region is simply the interval between the minimum and maximum values. In multivariate space, an interpolation region is defined by a convex hull which is the smallest convex area containing all the training data points. The simplest method to estimate a convex hull, which is also used in the present study, is taking the ranges of the individual feature values. The ranges define a multi-dimensional hyper-rectangle encompassing the convex hull.[166] The compounds with feature values outside the extremes of the training set are considered as outliers which are out of the applicability domain, and the prediction of activities of the outliers are considered unreliable.

Ensemble docking has been extensively investigated and employed in the virtual screening study effectively.[145] Herein, we employed the ensemble docking approach to further examine the predicted binding affinities of the hit compounds obtained from the hierarchical QSAR screening. The protocol and procedure of selecting and preparing the protein X-ray crystal structures for ensemble docking have been discussed in detail before. In the present work, 18 protein structures were selected, including 1Q3D, 1Q3W, 1Q41, 1Q4L, 1Q5K, 1R0E, 1UV5, 2O5K, 2OW3, 2JID, 3DU8, 3F7Z, 3F88, 3GB2, 3I4B, 3L1S, 3M1S, and 3PUP. The ensemble docking was performed using the Glide XP[119] scoring function implemented in the Schrödinger Suite 2010.

The final virtual screening protocol is illustrated in Figure 4.2. The screening chemical database is the KINASet collection of more than 12,000 drug-like small molecules, rationally selected from ChemBridge's EXPRESS-Pick^{TM} collection via desirable chemical group filtering and 3D pharmacophore query. The compounds in the chemical database underwent the same process of structural minimization and partial charge calculation. The hit compounds with high predicted pIC$_{50}$ values and reasonable docking scores were selected and subjected to the GSK-3β inhibition bioassays as described before.[145] The same reference compound, indirubin-3'-monoxime with reported IC$_{50}$ value as 190 nM, was employed to verify the reliability of the bioassay protocol.

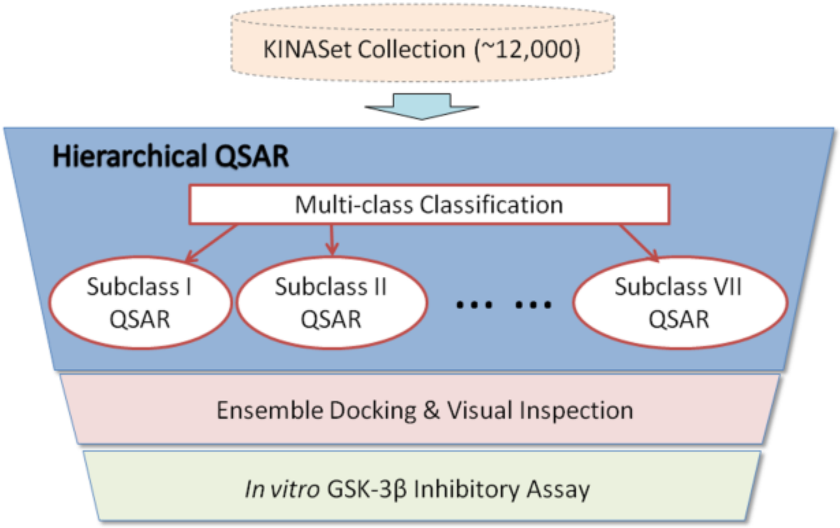

KINASet Collection (~12,000)
Hierarchical QSAR
Multi-class Classification
Subclass I QSAR
Subclass II QSAR
… …
Subclass VII QSAR
Ensemble Docking & Visual Inspection
In vitro GSK-3β Inhibitory Assay

Figure 4.2. The flowchart for high throughput virtual screening performed on the KINASet collection of more than 12,000 drug-like molecules.

4.3. RESULTS AND DISCUSSIONS

4.3.1. REGRESSION MODELS

The statistical results for all the regression models are summarized in Table 4.5. In order to systematically investigate the two cutting-edge machine learning algorithms (SVM and RF) in combination with feature selection, four models were built upon each group of compounds and the prediction abilities were validated both internally, using LOO and 5-fold cross-validation, and externally using an independent test set. The PSA-SVM model employed the SVM algorithm combined with the PS algorithm as feature selection; the RBE-RF model employed the RF algorithm with the RBE feature selection to remove the less important features. In general, the SVM models without any feature selection were over-fitted to the training data and the over-fitting problems are demonstrated by their low predictive ability (Table 4.5). For the worst case scenario, although the training R^2 of the SVM models built upon the Group I and Group III data sets is ~0.9, the predictive R^2 for both cross-validation and independent test sets are < 0.1.

However, by implementing feature selection the predictive power of the PSA-SVM models was significantly boosted. The majority of the models with the highest predictive R^2 for both cross-validation and independent test sets were produced using the PSA-SVM method. In contrast to SVM, RF is typically insensitive to the impact of redundant and irrelevant descriptors. This is supported by the high predictive R^2 for the independent test sets, most of which are comparable to the ones obtained by the PSA-SVM method. Despite the good performance in external validation, the internal validations using cross-validation yielded relatively low predictions accuracies. The majority of the models built by RF without feature selection have cross-validated $R^2 < 0.5$. The RBE feature selection can significantly boost the performance on internal validation while still retaining good predictive ability in external validation. In order to determine how many important descriptors are necessary for the RF models, the internal validation using 100 replications of 5-fold cross-validation was performed and the best median value was used to indicate the optimal size of the feature subset (Figure 4.3).

Table 4.5. The statistical results for the QSAR regression models.

Models[a]	Training Set			Cross Validation				Test Set		
	R^2_{train}	$RMSE_{train}$	F_{train}	R^2_{LOO}	$RMSE_{LOO}$	R^2_{CV}	$RMSE_{CV}$	R^2_{test}	$RMSE_{test}$	F_{test}
Group I										
PSA-SVM	0.997	0.006	29639	0.657	0.592	0.601	0.693	0.585	0.424	85.2
SVM	0.991	0.046	4849.6	0.093	1.802	0.059	1.793	0.026	0.990	6.5
RBE-RF	0.947	0.132	1516.3	0.533	0.810	0.526	0.828	0.734	0.248	111.8
RF	0.948	0.143	1325.5	0.464	0.928	0.427	0.992	0.748	0.249	101.7
Group II										
PSA-SVM	0.988	0.010	8685.1	0.798	0.150	0.768	0.172	0.828	0.144	149.5
SVM	0.991	0.013	5753.6	0.259	0.556	0.242	0.571	0.468	0.324	10.9
RBE-RF	0.945	0.064	923.6	0.471	0.392	0.499	0.374	0.768	0.145	53.3
RF	0.950	0.063	915.7	0.429	0.424	0.411	0.437	0.784	0.137	56.8
Group III										
PSA-SVM	0.998	0.004	31892	0.713	0.486	0.623	0.628	0.611	0.488	14.0
SVM	0.989	0.067	1192.2	0.099	1.671	0.047	1.673	0.011	1.311	2.1
RBE-RF	0.926	0.181	391.0	0.384	1.005	0.396	0.991	0.105	1.211	6.4
RF	0.953	0.179	349.1	0.260	1.209	0.231	1.256	0.224	1.004	6.4
Group IV										
PSA-SVM	0.921	0.092	738.2	0.613	0.465	0.551	0.558	0.641	0.521	46.5
SVM	0.999	0.002	31060	0.253	0.876	0.215	0.919	0.513	0.722	4.9
RBE-RF	0.952	0.109	400.8	0.499	0.614	0.433	0.685	0.680	0.447	11.8
RF	0.964	0.136	268.7	0.258	0.883	0.210	0.932	0.763	0.415	11.2
Group V										

PSA-SVM	0.986	0.017	2749.9	0.869	0.142	0.851	0.164	0.771	0.168	43.4
SVM	1.000	0.001	81620	0.231	0.833	0.214	0.853	0.821	0.104	27.3
RBE-RF	0.954	0.061	624.5	0.654	0.371	0.660	0.366	0.719	0.157	36.8
RF	0.949	0.074	479.8	0.568	0.467	0.550	0.490	0.750	0.129	38.1
Group VI										
PSA-SVM	0.991	0.006	4631.7	0.689	0.205	0.594	0.276	0.846	0.081	64.7
SVM	0.997	0.005	5294.5	0.198	0.545	0.148	0.569	0.586	0.268	1.8
RBE-RF	0.846	0.112	173.7	0.406	0.401	0.561	0.290	0.679	0.180	22.4
RF	0.945	0.057	343.3	0.413	0.387	0.399	0.397	0.676	0.148	17.4
Group VII										
PSA-SVM	0.998	0.004	18282.0	0.838	0.323	0.836	0.328	0.681	0.513	24.9
SVM	1.000	0.000	295840.0	0.416	1.179	0.379	1.244	0.546	0.839	2.0
RBE-RF	0.931	0.148	402.2	0.674	0.604	0.679	0.595	0.198	1.174	5.4
RF	0.959	0.128	418.4	0.541	0.866	0.521	0.917	0.614	0.588	7.9

[a] The selected models are highlighted by bold.

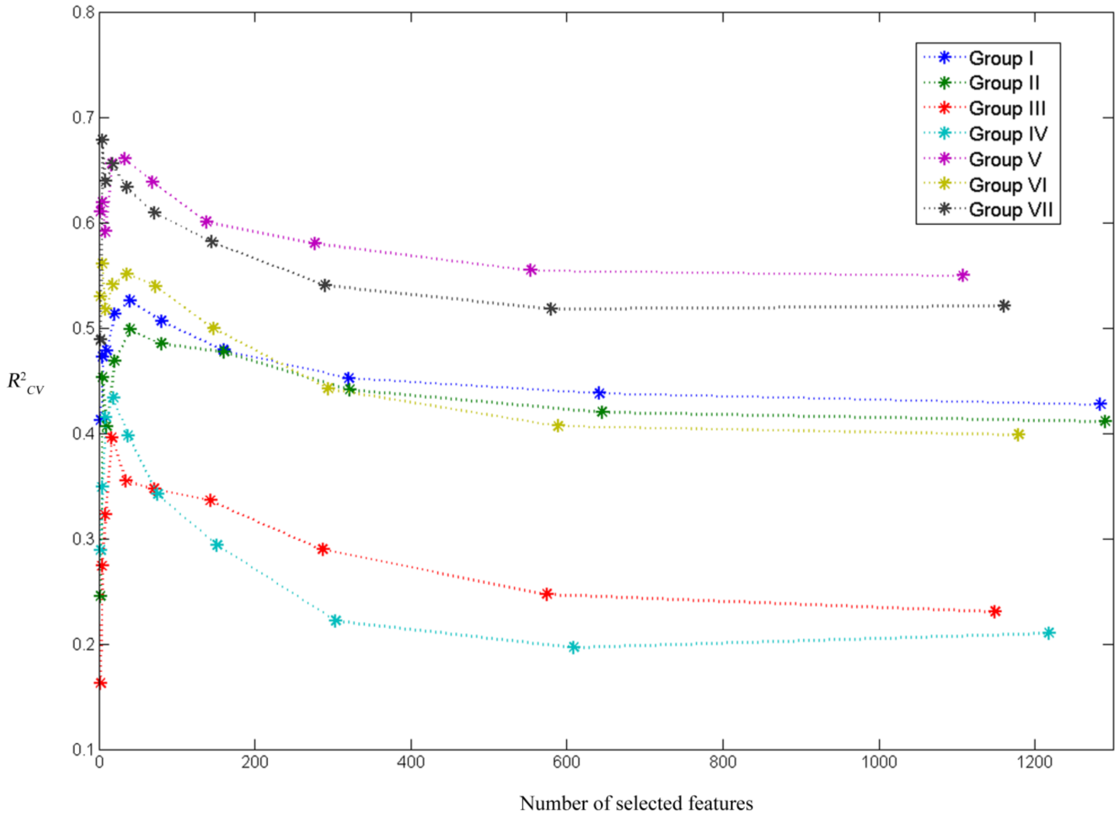


Figure 4.3. The number of features selected by RBE-RF and the internal validation R^2_{CV} (the median value from 100 replications of 5-fold cross-validation on the training set) for the various groups of compounds. The optimal number of features can be determined from the maximum R^2_{CV}.

The lack of correlation between the internal validation and external validation has been established earlier,[149] and also has been demonstrated in the present study. Evidence showing this lack of correlation for the RF algorithm is found for the majority of the models. In the context of the PSA-SVM models, the model for Group VI has the best performance on external validation ($R^2{}_{test}$=0.846, $RMSE_{test}$=0.081, F_{test}=64.7), but relatively poor performance on internal validation ($R^2{}_{LOO}$=0.689, $RMSE_{LOO}$=0.205, $R^2{}_{CV}$=0.594, $RMSE_{CV}$=0.276). However, the model for Group VII has the highest performance on internal validation ($R^2{}_{LOO}$=0.838, $RMSE_{LOO}$=0.323, $R^2{}_{CV}$=0.836, $RMSE_{CV}$=0.328), but relatively low prediction for external validation ($R^2{}_{test}$=0.681, $RMSE_{test}$=0.513, F_{test}=24.9). Nevertheless, the low correlation is not always found. By using the PSA-SVM method, the models for Group II and Group V performed consistently well on both internal and external validation, and the models for Group III and Group IV have consistently moderate predictions. Consequently, in order to make a reliable prediction, a set of criteria for both internal and external validation was formulated in the present study for model selection, namely: LOO cross-validated $R^2 > 0.5$, median value for 100 replications of 5-fold cross-validated $R^2 > 0.5$, and predictive $R^2 > 0.6$ for the independent test set.

Based on the criteria, we selected the best predictive regression models for each group of compounds. For the Group I data set, the best predictive model is RBE-RF characterized by R^2_{LOO}=0.533, $RMSE_{LOO}$=0.810, R^2_{CV}=0.526, $RMSE_{CV}$=0.828, R^2_{test}=0.734, $RMSE_{test}$=0.248, and F_{test}=111.8. This is the only group of compounds for which the RBE-RF method was able to produce a reliable prediction. The PSA-SVM method which was employed for the rest of the groups of compounds during the model selection cannot be used for the Group I data set, since the predictive R^2 for the independent test set is < 0.6 even though the cross-validated R^2 is > 0.6. This suggests that the RBE-RF algorithm compared to the PSA-SVM algorithm is less likely to be over-fitted, especially when it is used to deal with very large data sets consisting of structurally diverse compounds. The over-fitting problem of the PSA-SVM algorithm in certain cases may be attributed to the fact that the prediction performance of the selected feature subset is evaluated by interval validation using 5-fold cross-validation, which is a necessary but insufficient estimation of the predictive power. However, this over-fitting problem was only observed in certain cases. In general, the PSA-SVM method performed fairly well compared to other methods investigated in the present work, and indeed, it produced the best predictive models for the other group of compounds.

The best predictive models for Group II and Group V have high prediction performances: R^2_{LOO}=0.798, $RMSE_{LOO}$=0.150, R^2_{CV}=0.768, $RMSE_{CV}$=0.172, R^2_{test}=0.828, $RMSE_{test}$=0.144, and F_{test}=149.5 for Group II, R^2_{LOO}=0.869, $RMSE_{LOO}$=0.142, R^2_{CV}=0.851, $RMSE_{CV}$=0.164, R^2_{test}=0.771, $RMSE_{test}$=0.168, and F_{test}=43.4 for Group V. However, the best predictive models for Group III and Group IV have moderate prediction performances: R^2_{LOO}=0.713, $RMSE_{LOO}$=0.486, R^2_{CV}=0.623, $RMSE_{CV}$=0.628, R^2_{test}=0.611, $RMSE_{test}$=0.488, and F_{test}=14.0 for Group III, R^2_{LOO}=0.613, $RMSE_{LOO}$=0.465, R^2_{CV}=0.551, $RMSE_{CV}$=0.558, R^2_{test}=0.641, $RMSE_{test}$=0.521, and F_{test}=46.5 for Group IV. The moderate predictive power for the latter two groups of compounds may be caused by the Groups' intrinsic structural dissimilarity, which would translate into the lack of similar structural representations in the training set. It implies that the compounds in the training set cannot consistently reflect the influence of the structural modifications on the changes of the bioactivities.

The final PSA-SVM models were refined by optimizing the tuning parameters. The optimal tuning parameters and the number of selected features are tabulated in Table 4.6. The final RBE-RF models were established based on the default tuning parameters, and the optimal number of features were determined as described above. The best predictive regression models were employed at the lower level of hierarchical QSAR model. The selected features and their descriptions for the RBE-RF model for Group I and optimized PSA-SVM models for Group II through VII are provided in the supporting information (Appendix: J−N). In addition to the internal and external validation, the Y-randomization test described above was carried out to ensure that the best predictive regression models did not merely capture noise. As expected, all the models built based on the randomized bioactivities yielded fairly low predictive R^2 which was always < 0.1 for both LOO cross-validation and the independent test set (Table 4.7). These results further confirmed the robustness and reliability of the selected regression models which uncovered legitimate correlations between the molecular descriptors and the biological activities. The strong correlations can be used to predict the bioactivities of structurally novel compounds as GSK-3 inhibitors.

Table 4.6. The number of selected features and the optimal tuning parameters for PSA-SVM and SVM models.

Group	Model	No. of feat.	*C*	*γ*	*ε*
		multi-class classification model			
Whole Group	PSA-SVM	10	40	0.060	–
	SVM	1357	42	0.010	–
		regression models			
Group I	PSA-SVM	25	60	0.054	0.08
	SVM	1284	52	0.020	0.224
Group II	PSA-SVM	20	50	0.036	0.102
	SVM	1299	48	0.004	0.124
Group III	PSA-SVM	18	56	0.068	0.060
	SVM	1149	56	0.132	0.268
Group IV	PSA-SVM	20	20	0.038	0.018
	SVM	1218	40	0.004	0.048
Group V	PSA-SVM	9	48	0.100	0.112
	SVM	1108	20	0.004	0.024
Group VI	PSA-SVM	12	48	0.064	0.080
	SVM	1179	32	0.004	0.072
Group VII	PSA-SVM	11	40	0.064	0.060
	SVM	1161	32	0.004	0.016

Table 4.7. The statistical results for Y-randomization test.

Group	R^2_{LOO} [a]	$RMSE_{LOO}$ [a]	R^2_{test} [a]	$RMSE_{test}$ [a]
Group I[b]	0.009	2.423	0.021	1.253
	(0.014)	(0.247)	(0.032)	(0.148)
Group II[c]	0.011	1.420	0.038	1.307
	(0.015)	(0.195)	(0.049)	(0.338)
Group III[c]	0.024	2.923	0.063	2.960
	(0.029)	(0.565)	(0.082)	(1.127)
Group IV[c]	0.034	1.916	0.057	2.084
	(0.043)	(0.371)	(0.083)	(0.534)
Group V[c]	0.036	2.157	0.055	1.709
	(0.043)	(0.571)	(0.065)	(0.645)
Group VI[c]	0.033	1.357	0.092	1.218
	(0.046)	(0.360)	(0.112)	(0.479)
Group VII[c]	0.050	3.457	0.088	3.578
	(0.071)	(0.976)	(0.115)	(1.266)

[a] The number represents the mean values of 100 replications. The number in the parentheses represents the standard deviation of 100 replications. [b] Model III based on RBE-RF method. [c] Model I based on PSA-SVM method.

4.3.2. MULTI-CLASS CLASSIFICATION MODELS

The statistical results for all the multi-class classification models are summarized in Table 4.8. Three out of four models yielded predictive accuracies >0.9 for both internal and external validations. The high prediction of a multi-class classification model is derived from the fact that the chemical structures of a given group of molecules are unique and differ from those of other groups of molecules, and such structural diversity can be well described by the DragonX descriptors. RBE-RF model, which produced perfect prediction for both training and test set of compounds, was employed at the upper level of the hierarchical QSAR model. 42 selected features based on RBE approach and their descriptions are summarized in Appendix: P.

Table 4.8. The statistical results for the QSAR classification models.

Model	Accuracy on training set	Cross-validation accuracy on training set		Accuracy on test set
		LOO	5-fold	
PSA-SVM	0.997	0.995	0.992	0.986
SVM	1.000	0.607	0.570	0.587
RBE-RF	1.000	0.988	0.992	1.000
RF	1.000	0.988	0.983	0.993

4.3.3. HIERARCHICAL QSAR AND VIRTUAL SCREENING

To evaluate the predicative performance of hierarchical QSAR model on the collection of 728 GSK-3β inhibitors, a predictive model with multi-level structure was employed in a top-down way. The same division of training and test set was made, which lead to the training set with 587 compounds and test set with 141 compounds. Since the selected multi-class classification model (RBE-RF model) at the upper level yielded perfect prediction for all of the compounds in both training and test set, the predicted pIC_{50} values can be obtained from the best predictive regression models selected for each group of compounds at the lower level (Figure 4.4). The high predictive ability is indicated by the following statistical results: R^2_{train}=0.967, $RMSE_{train}$=0.054, F_{train}=13630, R^2_{test}=0.752, $RMSE_{test}$=0.265, and F_{test}=537.9.

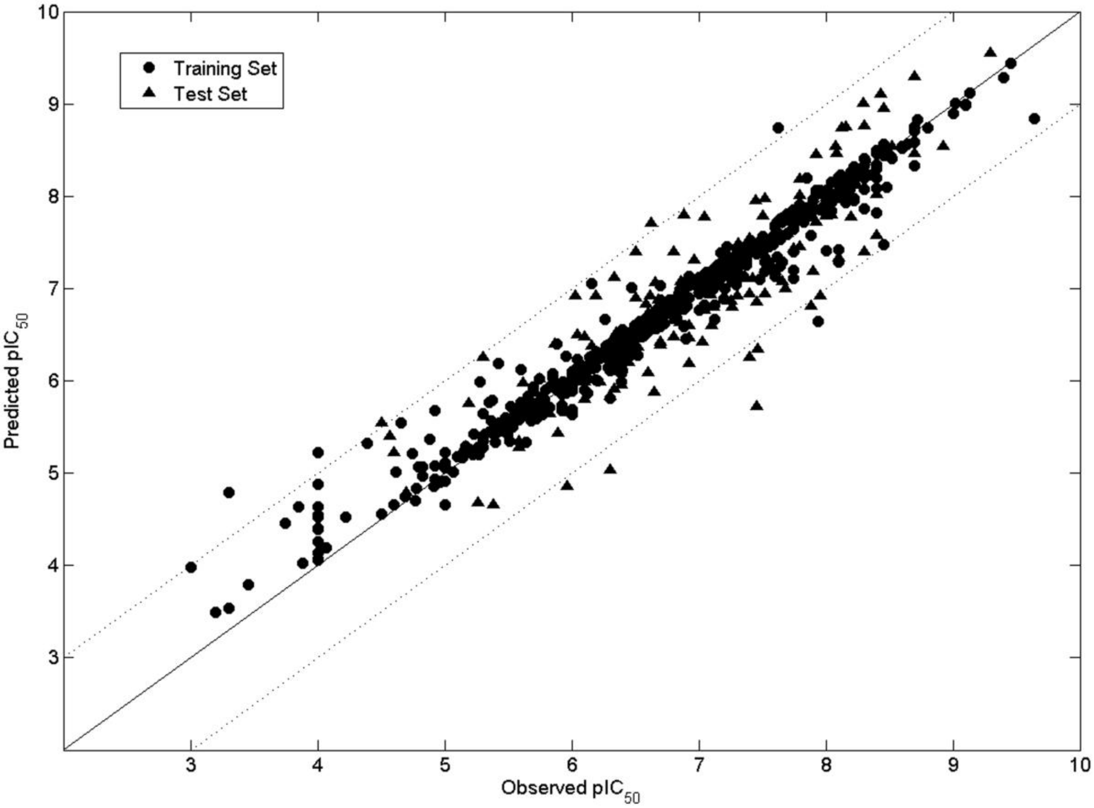
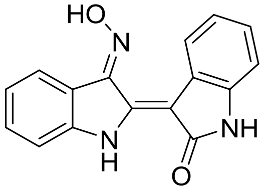
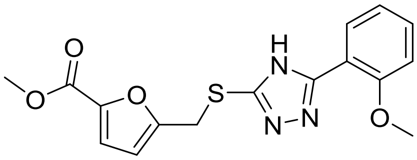
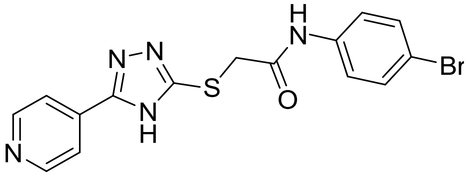
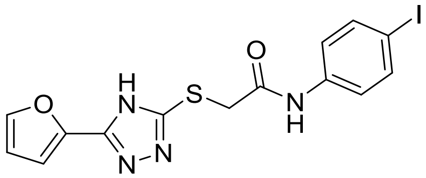
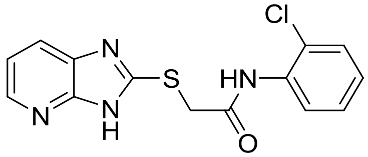
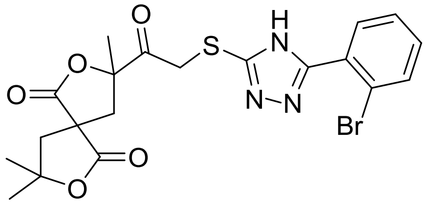


Figure 4.4. Correlation of observed versus predicted pIC_{50} values based on hierarchical QSAR model for the whole collection of 728 compounds. Circles represent the compounds in the training set and triangles represent the compounds in the test set.

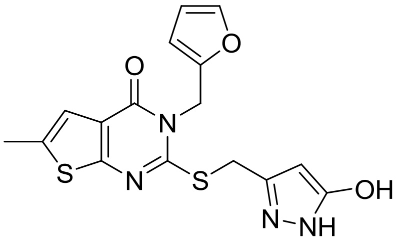
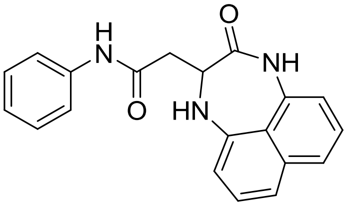
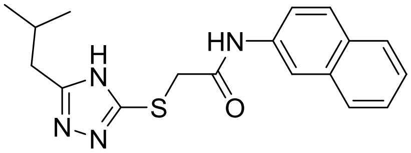
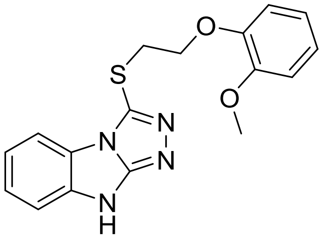
To examine the predictive ability of the hierarchical QSAR in the mixed ligand/structure-based virtual screening protocol from the experimental perspective, 5 compounds with high predicted pIC_{50} values (> 8.0) and acceptable docking scores were purchased from ChemBridge Corp., and the GSK-3β inhibitory activities of these compounds were tested according to the experiments described before. Their chemical structures and biological activities are listed in Table 4.9. Two compounds consisting of the same 1,2,4-triazole-3-thiol scaffold exhibit inhibition >50% at 10 μM concentration and low micromolar IC_{50} values. The new scaffold, which has not been identified before, can be employed to facilitate the discovery of structurally novel GSK-3β inhibitors.

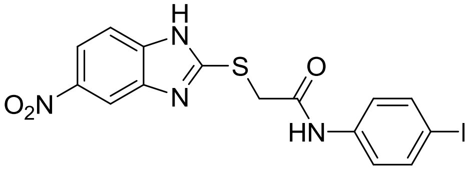
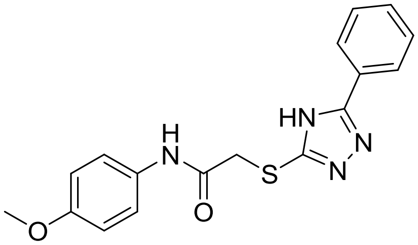
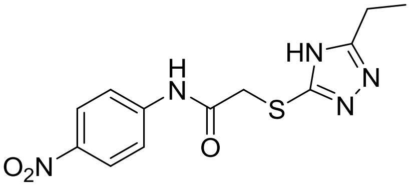
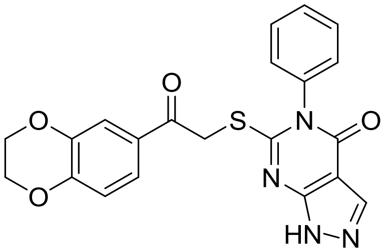
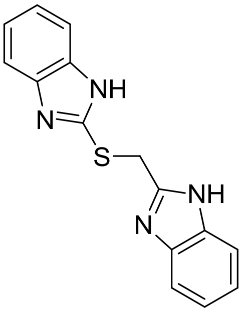
Table 4.9. Biological activities of the hit compounds.

ID	Chemical structure	QSAR	Docking Score	%inhibition (10 μM)	Ki (μM)	IC_{50} (μM)
Ref.				97%	0.07	0.13
1		8.526 (Group VI)	−9.788 (3I4B)	10%		
2		8.647 (Group II)	−9.445 (3DU8)	73%	3.53	7.05
3		8.475 (Group II)	−9.090 (3F88)	64%	2.69	5.38
4		8.610 (Group II)	−8.911 (3M1S)	23%		
5		8.854 (Group V)	−8.203 (2JLD)	22%		

We also purchased and tested the inhibitory activities of 9 potential hit compounds with moderate predicted bioactivities (predicted pIC_{50} from 6 to 8). However, they are typically false positive hits with inhibitory percentage < 50% at 10 μM (Table 4.10), and we did not bioassay them for the exact IC_{50} values. The low prediction for those false positive hits can be attributed to the 3D conformations of those compounds, which did not undergo conformational search for global minimum and may be fairly different from the possible binding conformation. The 3D conformation may have impact on the QSAR predictions since a few 3D descriptors employed in the models are conformational dependent.

Table 4.10. Biological activities of the false positive hits.

ID	Chemical structure	QSAR	Docking Score	%inhibition (10 μM)
1		6.976 (Group I)	-10.068 (3PUP)	12%
2		6.899 (Group I)	-10.025 (3L1S)	39%
3		7.549 (Group III)	-9.996 (1Q5K)	32%
4		7.118 (Group III)	-9.454 (3L1S)	11%

5		7.336 (Group V)	-9.436 (1Q5K)	8%
6		6.939 (Group II)	-9.370 (3L1S)	16%
7		7.849 (Group VI)	-8.808 (3F88)	6%
8		7.744 (Group VI)	-8.431 (1R0E)	14%
9		7.262 (Group V)	-8.248 (1Q5K)	29%

4.4. CONCLUSIONS

In order to identify novel GSK-3 inhibitors, we successfully implemented a mixed ligand/structure-based virtual screening protocol. In terms of the ligand-based approach, we constructed a hierarchical QSAR model which adopts a multi-level structure applied to integrate and analyze complex data sets from multiple experimental sources. In terms of the structure-based approach, we employed ensemble docking as an effective approach to evaluate the predicted binding affinities. The hit compounds obtained from virtual screening underwent experimental validation. The bioassay results showed that 2 out of 5 hit compounds are indeed GSK-3β inhibitors, exhibiting low micromolar activities. To build highly predicative classification and regression models in the multi-level structure, four different methods involving SVM and RF with or without feature selection were explored. The best regression models for the lower level were selected based on both internal and external validations, and the best classification model at the upper level yielded perfect prediction for the compounds in both training and set. The significance of feature selection in building predictive learning models was investigated, and there is no guarantee that a single approach can generate the best prediction for all the data set. Hence, systematic studies on different approaches are required to produce the best prediction for a particular data set. Based on the overall predictive performance and the successful virtual screening study, we can conclude that the proposed hierarchical QSAR model which makes predictions based on the most reliable models constructed using structurally similar compounds can be employed as an effective approach in virtual screening experiments.

Chapter 5. IMPLEMENTATION OF MULTIPLE-INSTANCE LEARNING IN DRUG ACTIVITY PREDICTION

Gang Fu, Xiaofei Nan, Haining Liu, Ronak Y. Patel, Pankaj R. Daga, Yixin Chen, Dawn E. Wilkins, Robert J. Doerksen

5.1. INTRODUCTION

In the context of molecular modeling and drug discovery research, it is imperative to specify which conformations of a given molecule are responsible for the observed biological activity. Due to structural flexibility, a molecule may adopt a wide range of conformers and the identification of the bioactive conformers is extremely important in order to understand the recognition mechanism between small molecules and proteins, which is crucial in drug discovery and development. Until now, the most reliable approach to obtain the bioactive conformer is to use the X-ray crystal structure of a ligand-protein complex; however, the number of such structures is limited because of the experimental difficulty in obtaining the crystals, especially for transmembrane proteins. We were interested to apply to this problem a machine-learning approach which does not require crystal structures, named multiple-instance learning (MIL) via embedded instance selection (MILES). MILES has been demonstrated as an efficient and accurate approach to solve different multiple-instance problems.[167] In the context of drug activity prediction, MILES enables the construction of a quantitative structure-activity relationship (QSAR) model, and subsequently the identification of bioactive conformers.

In the context of drug activity prediction, the observed biological activity is associated with a single molecule (bag) without knowing which conformer or conformers (instances) are responsible. Furthermore, a molecule is biologically active if and only if at least one of its conformers is responsible for the observed bioactivity; and the molecule is inactive if none of its conformers is responsible (Figure 5.1). A difficulty in implementation arises from the fact that different molecules have a different number of conformers, since some molecules having multiple rotatable bonds are highly flexible and others with rigid structures only have a small numbers of conformers.

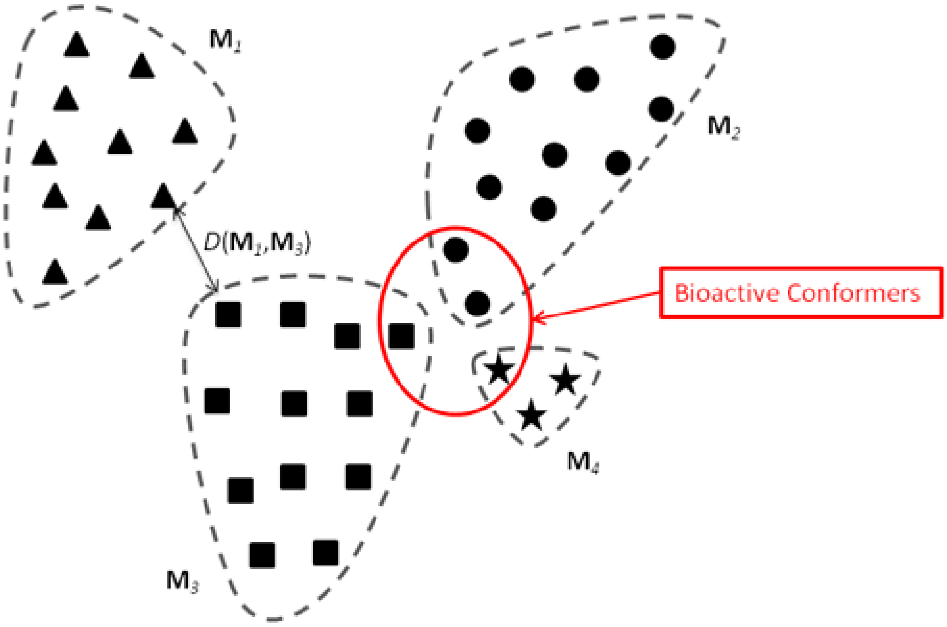
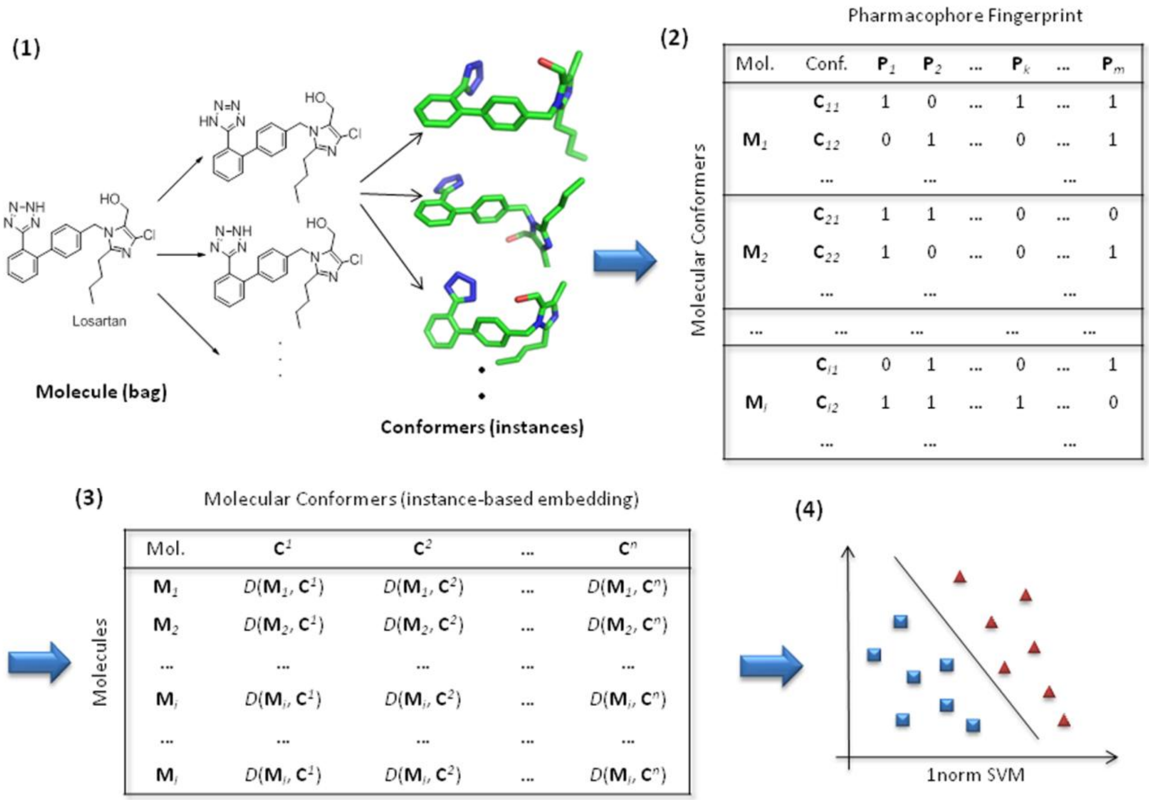


Figure 5.1. Cartoon representation of the relationship between molecules and conformers. Mi, i=1, 2, 3, 4 represent the molecules (bags), circled by dashed lines. The solid triangles in M1, circles in M2, squares in M3, and stars in M4 represent conformers for different molecules. Molecules 2, 3, and 4 were biologically active since they had at least one bioactive conformer, whereas molecule 1 was inactive since none of its conformers was bioactive. The distance between two molecules, M1 and M3, was calculated by the minimum distance D(M1, M3).

The overall strategy for structural and data mining using MILES (Figure 5.2) is summarized here. First of all, a complete sampling of conformational space provides a large number of conformers for each molecule. The molecules are themselves each already labelled as either positive or negative. However, the labels for the conformers are unavailable during the model generation. Each conformer is denoted by a unique pharmacophore fingerprint which is a superior feature-based 3D descriptor unveiling structural similarity and diversity.[168-171] The pharmacophore fingerprint is encoded into a binary string which indicates the presence or absence of a match to individual pharmacophore models. Since the exhaustively enumerated fingerprints have millions of bits, which may be beyond computational limits, a significance analysis of pharmacophore models[172] is employed to determine the optimal subset of bits of the fingerprint. Subsequently, MILES converts the MIL to a standard supervised learning problem by embedding bags (molecules) into an instance-based (conformer-based) feature space via structural dissimilarity measures.[173] Finally, 1-norm SVM is applied to select the most important features, identifying the highly significant conformers which help the most to distinguish active and inactive molecules, and, simultaneously, to construct a predictive classification model.



Mol.	Conf.	P_1	P_2	...	P_k	...	P_m
M_1	C_{11}	1	0	...	1	...	1
	C_{12}	0	1	...	0	...	1
	...		...			...	
M_2	C_{21}	1	1	...	0	...	0
	C_{22}	1	0	...	0	...	1
	...		...			...	
...	...		...		...		...
M_i	C_{i1}	0	1	...	0	...	1
	C_{i2}	1	1	...	1	...	0
	...		...			...	

Mol.	C^1	C^2	...	C^n
M_1	$D(M_1, C^1)$	$D(M_1, C^2)$	...	$D(M_1, C^n)$
M_2	$D(M_2, C^1)$	$D(M_2, C^2)$	...	$D(M_2, C^n)$
...	...	...	...	...
M_i	$D(M_i, C^1)$	$D(M_i, C^2)$	...	$D(M_i, C^n)$
...	...	...	...	...
M_l	$D(M_l, C^1)$	$D(M_l, C^2)$	...	$D(M_l, C^n)$

Figure 5.2. Overview of the MILES approach: (1) Structure preprocessing and conformational sampling. (2) Creating pharmacophore fingerprints and significance analysis of pharmacophore models. (3) Instance-based feature mapping based on structural similarity measures. (4) Joint feature selection and classification using 1-norm SVM.

In the present work, MILES has been applied to study the biological activities of two sets of molecules: GSK-3 inhibition data set and GSK-3/CDK1 selective inhibition data set. The first data set was explored to identify conformers significant for potent GSK-3 binding affinity; and the second data set was explored to investigate the conformers contributing to GSK-3/CDK1 binding selectivity. Based on our calculations, MILES is highly competitive with the classical QSAR approaches which do not include instance-based feature mapping in terms of predictive abilities. Meanwhile, MILES has been validated as a useful approach to identify the bioactive conformers, which contribute to the classification of active and inactive molecules.

5.2. METHODS

5.2.1. DATA SET PREPARATION

Two data sets were compiled through extensive literature search. Data set **I** included all molecules exhibiting inhibitory activities for GSK-3. Data set **II** included the compounds with reported inhibitory activities for both GSK-3 and CDK1. The datasets are publicly available (http://pars.cs.olemiss.edu/GangFu/MILES-project). The molecules collected for each data set were labelled as either positive or negative. A positive molecule either has a high binding affinity with GSK-3 in data set **I**, or has selective binding preference toward GSK-3 rather than CDK1 in data set **II**; whereas a negative molecule either has a low binding affinity with GSK-3, or does not exhibit selectivity toward GSK-3 over CDK1. A single cutoff value has been widely used in the development of classification models. However, it is inaccurate to use a single cutoff value for the separation of continuous biological activities in the context of drug activity prediction. The biological activities are represented by continuous numbers, and the small differences between the values above and below the cutoff value cannot imply the distinct nature of binding affinity. Furthermore, the small difference in the bioassay results may arise from systematic errors introduced by different experimental protocols used in different labs, so it cannot be used as solid evidence for the classification of molecules. Therefore, multiple cutoff values were employed to separate molecules into positive and negative classes. For data set **I**, the molecules were categorized into positive and negative molecules using cutoff values of $IC_{50} \leq 50$ nM and $IC_{50} \geq 500$ nM, respectively. The molecules having inhibitory activities between the two cutoff values were considered as moderately active molecules, and were discarded from the data set. The wide margin between the two cutoff values was used to account for the variances in biological assays. This resulted in the selection of 260 positive and 258 negative compounds. For data set **II**, the molecules were considered as selective (positive) for GSK-3 over CDK1 if the inhibitory activity for GSK-3 is 10 times more than that for CDK1 and the molecules were considered as nonselective (negative) if the inhibitory activity for GSK-3 is less than that for CDK1. The cutoff values yielded 97 selective and 134 nonselective compounds.

External validation was achieved using an independent test set. The split of the data set into training and test sets was carried out using Kohonen self-organizing maps (SOM) in Canvas 1.4.[174] The SOM is trained using unsupervised learning to produce a square 2D grid map from the high dimensional input space. Each grid cell (neuron) contains a cluster of structurally similar molecules defined by the input vectors. The SOM takes advantage of clustering capabilities so that the selected training set can represent the independent test set in terms of the input space and chemical domains. Molecular pharmacophore fingerprints were used to describe the relevant structural information of the molecules and were used as input variables to build the SOM. The grid size of the map depends on the number of molecules in the data set. For data sets **I**, the Kohonen maps built included 10×10 neurons and 500 epochs. For the data set **II**, a

Kohonen map consisting of 8×8 neurons and 500 epochs was built. The molecules were then stratified and sampled from each neuron to select the training and test set molecules.

5.2.2. PREPROCESSING AND CONFORMATIONAL SAMPLING

The molecules (bags) can be represented by $\mathbf{M}_i$, $i=1,\cdots,l$ where l is the total number of molecules. The 3D molecular structures were generated using the Ligprep module from Schrödinger Suite 2011, and then subjected to preprocessing to enumerate all the possible tautomers. The protonation states of ionizable groups were set to match pH = 7.4, and the stereochemistry was retained from the original 3D structures. In order to explore the conformational space exhaustively, the mixed torsional/low mode sampling method was employed, using MacroModel from Schrödinger Suite 2011. The torsional sampling involves multiple Monte Carlo minimum searches for global exploration, and the low mode conformational search allows for automatic local exploration. The torsional increment for each rotatable bond was set to 15° and the maximum number of total steps for torsional sampling was 1,000. The energy window for saving structures was set to 83.7 kJ/mol (20 kcal/mol). The small torsional increment and wide energy window were employed to provide a reasonable coverage of the conformational space. Each enumerated conformer was energy minimized using the Powell-Reeves conjugate gradient method with default setup. To remove redundant conformations, the maximum atom deviation cutoff was set to 1.5 Å. So each molecule $\mathbf{M}_i$ has several possible conformers $\mathbf{C}_{ij}$, $j=1,\cdots,n_i$, where n_i is the number of conformers (instances) for molecule i.

In order to validate that MILES can identify the bioactive conformers, we seeded 12 co-crystallized conformers, one for each of 12 molecules, in the set of sampled conformers for data set **I**. The validation process will be described in the following sections.

5.2.3. GENERATION OF PHARMACOPHORE FINGERPRINTS

The pharmacophore fingerprint as a measure of molecular similarity and diversity based on 3D pharmacophoric shape was enumerated using Canvas 1.4. Each pharmacophore fingerprint associated with a unique conformer can be represented by a binary string, such as $\mathbf{P}_{ij} = \{\mathbf{p}_1,\cdots,\mathbf{p}_k, \cdots,\mathbf{p}_m\}$ and encodes quantitative structural information for conformer $\mathbf{C}_{ij}$, where each bit value $\mathbf{p}_k$, k=1,$\cdots$,m indicates the presence or absence of a single pharmacophore model, representing a unique 3D arrangement of a number of pharmacophore features. If the conformer fits the pharmacophore model for a particular k, in other words the functional groups of the conformer fully overlap on all the pharmacophore features in the model, $\mathbf{p}_k$ equals 1; otherwise, $\mathbf{p}_k$ equals 0. As a result, each conformer is associated with a unique pharmacophore fingerprint as a conformational signature, which enables us to describe quantitatively the 3D structural information. In the present study, only four-feature based models were employed in order to allow a reasonable description of 3D orientation of the structures and retain information about molecular chirality, which is lost in three-feature based models. The pharmacophore features employed in the models consist of hydrogen bond donor (D), hydrogen bond acceptor (A), hydrophobic group (H), negatively charged group (N), positively charged group (P), and aromatic ring (A). The maximum distance between pharmacophore features was set to 20.0 Å in order to be able to cover the largest molecular structures in the databases. The originally enumerated fingerprints were subject to filtering to remove the pharmacophore models present in less than 5% of the total number of molecules, since the pharmacophore models with a very low occurrence are not useful for discriminating between positive and negative classes.

5.2.4. SIGNIFICANCE ANALYSIS OF PHARMACOPHORE MODELS

The post-filtered pharmacophore fingerprints still have too many bits that lack information content, as indicated by too many ‘0’ values. Therefore a nonparametric supervised learning approach, motivated by the significance analysis of microarrays (SAM) algorithm proposed by Tibshirani *et al.*,[175] was applied to elucidate a consistent pattern from the numerous bits of pharmacophore fingerprints. The detailed implementation and customization of the relevant procedures has been described in.[172] The ranking score for each pharmacophore model was computed based on the occurrences of that model in each class, either positive or negative. That ranking score was then compared with a reference score computed from 500 random permutations of the class labels across all the molecules. If the difference between the true score and the reference score exceeds a cutoff threshold (called Δ) then that conveys statistical significance. The two-class t-statistic was used to estimate the percentile of truly significant pharmacophore models.

5.2.5. INSTANCE-BASED FEATURE MAPPING

MILES provides a framework to convert a MIL problem to a standard supervised learning problem via instance-based feature mapping. All the conformers (instances) belong to the instance-based feature space. For convenience, all conformers in all molecules were lined up together, and were re-indexed in the embedded feature space as $\mathbf{C}^r$, $r=1,\cdots,n$ where $n = \sum_{i=1}^{l} n_i$. Instance-based feature mapping can be accomplished using calculated structural dissimilarities. Different binary string distance measures were tested, including the Soergel distance, Dice distance, Manhattan distance, and Rogers-Tanimoto distance (Table 5.1). The range of each dissimilarity measure was normalized to be [0, 1] by definition. Since one molecule is defined as a bag of multiple conformers (instances), the dissimilarity measure for a molecule is calculated based on the minimum distance using the closest instance in the bag. The minimum distance calculation (Figure 5.1) extends the idea of the diverse density framework proposed for instance-based learning.[176]

Table 5.1. Metrics used for dissimilarity measurements

Dissimilarity Measure	Definition[a]
Soergel	$\frac{b+c}{a+b+c}$
Dice	$\frac{b+c}{2a+b+c}$
Manhattan	$\frac{b+c}{a+b+c+d}$
Rogers-Tanimoto	$\frac{2\times(b+c)}{a+d+2\times(b+c)}$

[a] Let P1 and P2 be two pharmacophore fingerprints, a be the count of bits which are set to 1 in both P1 and P2, b be the count of bits which are set to 1 in P1 but not in P2, c be the count of bits which are set to 1 in P2 but not in P1, and d be the count of bits which are set to 0 in both P1 and P2. So a is called the number of total matches, b and c are called the number of single matches, and d is called the number of no matches.

5.2.6. JOINT FEATURE SELECTION AND CLASSIFICATION

Since the molecules in the training sets are highly flexible, instance-based embedding, which provides a framework to convert a MIL problem to a traditional supervised learning problem, may produce a very high dimensional feature space. But many features are redundant or irrelevant, and do not play an important role in the classification of molecules as positive or negative. So an efficient feature selection model is required for selection of an optimal subset of instance-based features. Considering its excellent performance in many applications,[177] the 1-norm SVM method was chosen as a joint approach to construct classifiers and to select important features simultaneously.

The features selected as important for the classification problem of interest are called *prototype conformers*. The plus or minus sign of the feature coefficient indicates the positive or negative contribution, respectively, of each prototype conformer to the putative bioactive conformers for each individual molecule.

5.2.7. IDENTIFICATION OF BIOACTIVE CONFORMERS

One appealing advantage of the MILES algorithm is that it can identify the most significant instances in a bag according to their contributions to the classification of that bag. In the context of drug activity prediction, we can identify the most significant conformers, called the bioactive conformers, for each molecule. The putative bioactive conformers are the conformers that contributed the most to the classification of positive and negative molecules. The identification of bioactive conformers can be accomplished by calculating the contribution of each conformer of a given molecule. The contribution can be calculated with the assistance of the prototype conformers. Typically, the contribution is the total sum of the weighted distance between each conformer of a given molecule and the closest prototype conformers.

In order to validate the ability of MILES to identify the bioactive conformers, the contributions for the 12 seeded conformers, which were taken directly from co-crystallized complex structures, were calculated and ranked among all the conformers sampled for those 12 molecules. The PDB codes of 12 co-crystallized structures were 1Q5K, 2O5K, 2OW3, 1Q3W, 1UV5, 1Q41, 1R0E, 3F7Z, 3GB2, 3L1S, 1Q4L, and 1Q3D.

5.2.8. CLASSICAL QSAR METHODS WITHOUT INSTANCE-BASED EMBEDDING

In order to examine the predictive performance of MILES, conventional classification approaches based on classical QSAR principles were tested for comparison. Without instance-based embedding, the feature space for classical QSAR studies is based on the optimal subsets of the fingerprints selected through significance analysis of the pharmacophore models. The decision tree was constructed using the ‘classregtree’ function implemented in Matlab R2011b. Gini’s diversity index was used for recursive partitioning, and the minimal number of molecules per tree leaf was set as 3 to terminate tree growing. The MILES model was built in the pharmacophore-based feature space. The ensemble learning method, random forests, developed by Leo Breiman [163], has been demonstrated as one of the most powerful tools available for data exploration [164]. The Matlab implementation (randomforest-matlab v0.02) was used with default parameters.

5.3. RESULTS AND DISCUSSION

5.3.1. DATA SET PREPARATION AND CONFORMATIONAL SAMPLING

According to the criteria used to label positive and negative molecules, the number of molecules in each of two classes was balanced for four data sets. Data set **I** has 266 positive compounds including 199 in training set and 67 in test set, as well as 258 negative compounds including 188 in training set and 70 in test set; data set **II** has 97 positive compounds including 76 in training set and 21 in test set, as well as 134 negative compounds including 100 in training set and 34 in test set. The stratified sampling divided data sets into training and test sets at ratios around 3:1.

The total number of conformers generated for data set **I** was 22,648, and the total number of conformers generated for data set II was 12,961. The feature space constructed through instance-based embedding only consisted of the instances from training bags, in other words, the conformers from the molecules in the training set. The molecules in the test set were not used in the construction of the instance-based feature space. So the number of instance-based features used for embedding molecules in data set **I** was 17,249, and the number of instance-based features used for embedding molecules in data set **II** was 9,972.

5.3.2. SIGNIFICANCE ANALYSIS OF PHARMACOPHORE MODELS

Millions of pharmacophore models were originally enumerated for each data set (1,872,521 for data set **I** and 1,378,584 for data set **II**). After occurrence-based filtering, only a small portion of the pharmacophore models was retained for each data set (243,721 for data set **I** and 253,192 for data set **II**).

Significance analysis was subsequently performed upon those retained pharmacophore models. The number of statistically significant pharmacophore models was computed at the 90th percentile among 500 permutations using the classical *t*-statistic. The threshold values were set to 100 equally spaced intervals from 0 to the largest difference between the ranking scores and reference scores. As the threshold value increases in a bottom-up manner, the number of falsely significant pharmacophore models decreases, and the number of truly significant models remains roughly constant. So the optimal threshold values (Δ^*=1.77 for data set **I** and Δ^*=2.17 for data set **II**) for each data set can be obtained when the number of falsely significant pharmacophore models drops to zero. Subsequently, the optimal subsets of the pharmacophore fingerprint bits were obtained (2,979 for data set **I** and 10,010 for data set **II**).

In the context of MIL, the optimal subsets of the binary strings were used to calculate the dissimilarity between two conformers for instance-based feature mapping. For the classical QSAR methods, the optimal subsets of the fingerprints were used as the 3D descriptors in the pharmacophore-based feature space for building classification models.

5.3.3. PREDICTIVE PERFORMANCE OF MILES AND CLASSICAL QSAR METHODS

In the MILES model, the only tuning parameter λ was determined by a grid search. Five replications of 5-fold cross-validation were performed to assess the classification accuracies at each point over a fixed grid which ranged from 2^{-8} to 2^{5} with exponential increment in base 2. The median values for the 5 replications were used to find the optimal tuning parameters. During the cross-validation, the instance-based feature space was dynamically defined, which means that the conformers from the molecules in the internal test set, after random split of the training set, were excluded from the feature space. As a result, the optimal tuning parameters as well as the number of prototype conformers were obtained for four dissimilarity measures (Table 5.2).

Table 5.2. Optimization of tuning parameter λ and predictive performance for four different dissimilarity measures used in MILES

Data set	Dissimilarity measure	Cross-validation[a]	λ	*n*[b]	Training set		Test set	
					Accu	MCC	Accu	MCC
I	Soergel	0.777	8.000	196	0.972	0.944	0.854	0.714
	Dice	0.761	4.400	165	0.979	0.959	0.825	0.653
	Manhattan[c]	0.803	4.400	130	0.941	0.881	0.861	0.725
	Rogers-Tanimoto	0.801	4.000	153	0.961	0.923	0.861	0.725
II	Soergel	0.700	0.001	95	0.949	0.902	0.909	0.818
	Dice	0.709	0.002	81	0.949	0.902	0.927	0.851
	Manhattan	0.810	0.016	40	0.943	0.892	0.855	0.699
	Rogers-Tanimoto[c]	0.830	0.350	29	0.892	0.805	0.891	0.786

[a] The median classification accuracy for 5 replications of 5-fold cross-validation; [b] the number of prototype conformers; [c] The model selected based on the number of prototype conformers.

Based on the internal validation, the classification accuracies were similar within each data set using four different dissimilarity measures. However, the numbers of prototype conformers selected were much different. The dissimilarity measure which yielded the smallest number of selected prototype conformers was chosen as the best MILES model and used later for comparison with classical QSAR models without instance-based embedding.

After finding the optimal λ, a MILES model was identified from the training set and applied to the test set. In addition to comparing classification accuracy, denoted as the proportion of correct predictions, Mathews Correlation Coefficient (MCC)[178] was also employed a complementary indicator for the predictive performance. MCC not only takes into account true positives and true negatives as classification accuracy does, but also false positives and false negatives. Thus it is considered as a balanced measure of the performance of binary classification (Figure 5.2). In accordance to classification accuracy and MCC, the performance of different dissimilarity measures was dataset-specific. For data set **I**, both the Manhattan and Rogers-Tanimoto distances were top-ranked and performed equally well on the test set, whereas on the training set, the Rogers-Tanimoto distance performed slightly better than the Manhattan distance. In addition, the results did not change after removing the 12 seeded conformers which were used for the validation of identifying bioactive conformers. For data set **II**, the Rogers-Tanimoto distance which selected the minimum number of prototype conformers did not yield best predictive performance. However, the predictive abilities across different dissimilarity measures were similar.

After comparing the predictive performance of different dissimilarity measures in the MILES model, the predictive performance of MILES models was compared with that of conventional classification approaches, which are based on classical QSAR principles without instance-based embedding (Table 5.3). To find the optimal λ for 1-norm SVM on the basis of classical QSAR principles, the same procedure was employed, which resulted in the minimal subset of the most important pharmacophore models. For data set I, optimal λ was 0.001 with 223 selected pharmacophore models; for data set II, optimal λ was 0.8 with 50 selected pharmacophore models.

Table 5.3. Predictive performance for different models

Data set	Methods	Training set		Test set	
		Accuracy	MCC	Accuracy	MCC
I	MILES[a]	0.941	0.881	0.861	0.725
	Decision tree	0.915	0.830	0.781	0.569
	1-norm SVM	1.000	1.000	0.832	0.668
	Random forest	0.995	0.990	0.891	0.783
II	**MILES**[b]	0.892	0.805	0.891	0.786
	Decision tree	0.909	0.821	0.764	0.567
	1-norm SVM	0.955	0.912	0.891	0.767
	Random forest	0.897	0.810	0.873	0.755

[a] Manhattan dissimilarity measure; [b] Rogers-Tanimoto dissimilarity measure.

For data set **I**, the 1-norm SVM without instance-based embedding overfit the training set, producing perfect prediction on the training set and poor prediction on the test set. However, MILES performed fairly well on both the training and test sets without overfitting. MILES performed much better than decision trees and slightly worse than random forests in terms of the predictive power on the test set. For data set **II**, MILES outperformed all of the classical QSAR methods. It is noteworthy that the dissimilarity measure (Rogers-Tanimoto) used in comparison yielded slight lower prediction compared to Soergel and Dice distances. MILES performed fairly well on both training and test sets without overfitting, and its predictive power was highly comparable with other conventional QSAR approaches. It was interesting that the classification accuracy and MCC provided the same indications again, even for the comparison of different QSAR approaches.

5.3.4. IDENTIFICATION OF BIOACTIVE CONFORMERS

After examining the predictive ability of MILES, we tested the ability of MILES in the pursuit of the bioactive conformers. We made use of 12 co-crystallized structures of GSK-3 with bound small molecules, which adopt bioactive conformers in the complex structures (Table 5.4). The direct comparison between the structures of the co-crystallized conformers and the ones from conformational sampling is difficult and sometimes impossible, since the conformational sampling plus structural minimization may not provide the exact same conformations found in the co-crystallized complex, due to the lack of protein environment in the conformational search process. So we adopted an indirect validation method. We seeded the 12 co-crystallized conformers in the set of sampled conformers generated through extensive exploration of conformational space. Then we calculated their contributions $f\left(\mathbf{C}_{ij^*}\right)$ to the classification of the relevant positive molecules as described above (Table 5.4).

Three out of 12 molecules are highly flexible, adopting more than 100 conformers. For these three, MILES only correctly predicted one co-crystallized conformer as the third most significant conformer contributing to the classification of the molecule named AR. It incorrectly predicted the other two co-crystallized conformers as irrelevant conformers in terms of the contribution to the classification of benzoimidazole-1 and maleimide.

Table 5.4. Validations on the prediction of bioactive conformers

ID[a]	Name[b]	PDB ID	Contribution[c]	Rank[d]	*n*[e]
23	AR	1Q5K	2.792	3	117
37	Benzoimidazole-1	2O5K	0	N.A.[f]	138
50	Jonjon-1	2OW3	2.827	6	38
59	LM-4	1Q3W	0.858	1	2
60	LM-5	1UV5	11.941	1	3
77	LM-29	1Q41	8.576	2	7
97	Maleimide	1R0E	0	N.A. [f]	121
98	OxaD-0	3F7Z	10.629	1	53
99	OxaD-00	3GB2	4.637	2	9
153	Pyzo-11	3L1S	10.371	1	11
198	RM-0	1Q4L	5.568	2	25
199	Staurosporine	1Q3D	22.359	1	5

[a] Molecule index in the data set; [b] molecular name in the data set; [c] contribution calculated based on the weighted distance; [d] the rank in the set of contributions; [e] the number of conformers for each molecule; [f] the rank cannot be determined and the conformer was predicted to be irrelevant to classification based on the MILES method.

But for the molecules adopting less than 100 conformers, which had relatively rigid structures, MILES correctly predicted all the co-crystallized conformers as significant conformers for the classification of positive molecules. Five co-crystallized conformers were predicted to be the most significant conformers, i.e., the bioactive conformers; three co-crystallized conformers were predicted to be the second most significant conformers; and one co-crystallized conformer was predicted to be the sixth most significant conformer, based on the calculations of $f(\mathbf{C}_{ij^*})$. So the pursuit of bioactive conformers is easy for relatively rigid molecules and relatively more difficult for the highly flexible ones.

5.4. CONCLUSIONS

We have successfully implemented a multiple-instance learning (MIL) framework, multiple-instance learning via embedded instance selection (MILES), for drug activity prediction. The molecules and relevant conformers were described using superior 3D descriptors, pharmacophore fingerprints, encoded as binary strings. The instance-based embedding was accomplished using dissimilarity measures designed for calculations on binary strings. The joint feature selection and classification was accomplished using a wrapper model based on 1-norm SVM. We have used the approach for the prediction of the labels of molecules interacting with four therapeutic targets, including GSK-3, CBrs, and P-gp. Based on the predictive performance, our proposed approach was highly competitive with conventional classification approaches based on classical QSAR principle. Subsequently, we have validated that the proposed approach is highly useful in the pursuit of bioactive conformers using a set of 12 GSK-3 crystal structures with bound inhibitors.

Chapter 6. FUTURE PLAN ON LEAD COMPOUNDS

The most active lead compound we identified through *in silico* virtual screening carries a phthalimide scaffold, and phthalimide compounds have been reported to have inhibitory activities against AGC family of protein kinases including AKT, PDK1, p70S6K, and ROCK kinases, which involve in proliferative and neurodegenerative disorders.[179] The compounds can be easily synthesized through the Scheme I, and more structural analogs can be synthesized through the Scheme II. The GSK-3 inhibition of phthalimide has not been reported before. Hence, extensive structure-activity relationship can be explored to find promising drug candidates with potent inhibitory activities.

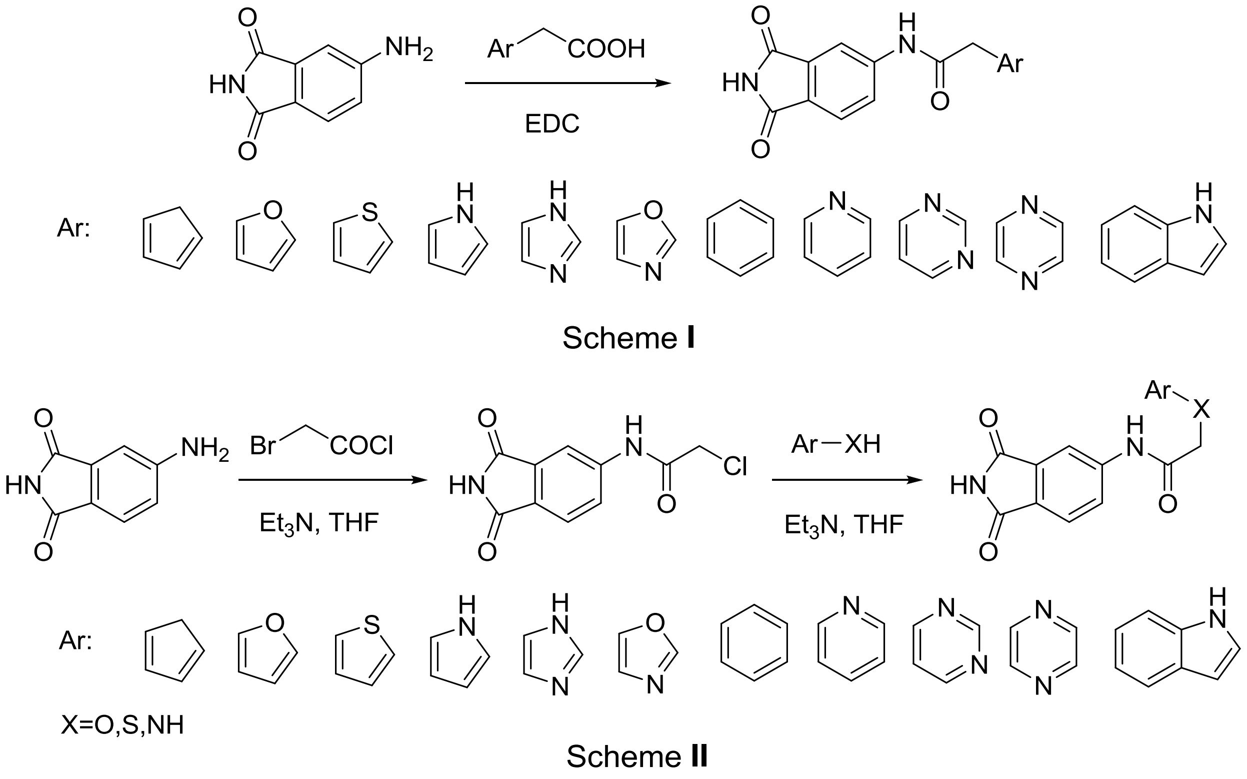


Figure 6.1. Synthetic route to modify chemical structure of lead compound.

BIBLIOGRAPHY

1. Doble, B. W.; Woodgett, J. R. GSK-3: tricks of the trade for a multi-tasking kinase. J. Cell Sci. 2003, 116, 1175-1186.

2. Cohen, P.; Goedert, M. GSK3 inhibitiors: development and therapeutic potential. Nat. Rev. Drug Discov. 2004, 3, 479-487.

3. Harwood, A. J. Regulation of GSK-3: a cellular multiprocessor. Cell 2001, 105, 821-824.

4. Seidensticker, M. J.; Behrens, J. Biochemical interactions in the wnt pathway. Biochim. Biophys. Acta 2000, 1495, 168-182.

5. Cohen, P.; Frame, S. The renaissance of GSK3. Nat. Rev. Mol. Cell Bio. 2001, 2, 769-776.

6. Jope, R. S.; Johnson, G. V. W. The glamour and gloom of glycogen synthase kinase-3. Trends Biochem. Sci. 2004, 29, 96-102.

7. Dajani, R.; Fraser, E.; Roe, S. M.; Young, N.; Good, V.; Dale, T. C.; Pearl, L. H. Crystal structure of glycogen synthase kinase 3beta: structural basis for phosphate-primed substrate specificity and autoinhibition. Cell 2001, 105, 721-732.

8. ter Haar, E.; Coll, J. T.; Austen, D. A.; Hsiao, H. M.; Swenson, L.; Jain, J. Structure of GSK3beta reveals a primed phosphorylation mechanism. Nat. Struct. Biol. 2001, 8, 593-596.

9. Frame, S.; Cohen, P.; Blondl, R. M. A common phosphate binding site explains the unique substrate specificity of GSK3 and its inactivation by phosphorylation. Mol. Cell. 2001, 7, 1321-1327.

10. Dajani, R.; Fraser, E.; Roe, S. M.; Yeo, M.; Good, V.; Thompson, V.; Dale, T. C.; Pearl, L. H. Structural basis for recruitment of glycogen synthase kinase 3beta to the axin-APC scaffold complex. EMBO J. 2003, 22, 494-501.

11. Bax, B.; Carter, P. S.; Lewis, C.; Guy, A. R.; Bridges, A.; Tanner, R.; Pettman, G.; Mannix, C.; Culbert, A. A.; Brown, M. J. B.; Smith, D. G.; Reith, A. D. The structure of phosphorylated GSK03beta complexed with a peptide, FRATtide, that inhibits beta-catenin phosphorylation. Structure 2001, 9, 1143-1152.

12. Woodgett, J. R. Molecular cloning and expression of glycogen synthase kinase-3/factor A. EMBO J. 1990, 9, 2431-2438.

13. Johnson, G. V.; Stoothoff, W. H. Tau phosphorylation in neuronal cell function and dysfuntion. J. Cell Sci. 2004, 117, 5721-5729.

14. Mukai, F.; Ishiguro, K.; Sano, Y.; Fujita, S. C. Alternative splicing isoform of tau protein kinase I/glycogen synthase kinase 3beta. J. Neurochem. 2002, 81, 1073-1083.

15. Drechsel, D. N.; Hyman, A. A.; Cobb, M. H.; Kirschner, M. W. Modulation of the dynamic instability of tubulin assembly by the microtubule-associated protein tau. Mol. Biol. Cell 1992, 3, 1141-1154.

16. Johnson, G. V.; Bailey, C. D. Tau, where are we now? J. Alzheimers Dis. 2002, 4, 375–398.

17. Hong, M.; Chen, D. C.; Klein, P. S.; Lee, V. M. Lithium reduces tau phosphorylation by inhibition of glycogen synthase kinase-3. J. Biol. Chem. 1997, 272, 25326–25332.

18. Lovestone, S. Alzheimer’s disease-like phosphorylation of the microtubule-associated protein tau by glycogen synthase kinase-3 in transfected mammalian cells. Curr. Biol. 1994, 4, 1077–1086.

19. Munoz Montano, J. R.; Moreno, F. J.; Avila, J.; Diaz Nido, J. Lithium inhibits Alzheimer’s disease-like tau protein phosphorylation in neurons. FEBS Lett. 1997, 411, 183–188.

20. Cheng, Y.; Zhang, Y.; McCammon, J. A. How does the cAMP-dependent protein kinase catalyze the phosphorylation reaction: an ab initio QM/MM study. J. Am. Chem. Soc. 2005, 127, 1553-1562.

21. Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Eyck, L. F. T. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. Proc. Natl. Acad. Sci. U.S.A. 2006, 103, 17783-17788.

22. Martinez, A.; Castro, A.; Dorronsoro, I.; Alonso, M. Glycogen synthase kinase 3 (GSK-3) inhibitors as new promising drugs for diabetes, neurodegeneration, cancer, and inflammation. Med. Res. Rev. 2002, 22, 373-384.

23. Martinez, A. Preclinical efficacy on GSK-3 inhibitors: towards a future generation of powerful drugs. Med. Res. Rev. 2008, 28, 773-769.

24. Alonso, M.; Martinez, A. GSK-3 inhibitors: Discoveries and developments. Curr. Med. Chem. 2004, 11, 755-763.

25. Medina, M.; Castro, A. Glycogen synthase kinase-3 (GSK-3) inhibitors reach the clinic. Curr. Opin. Drug Di. De. 2008, 11, 533-543.

26. Coghlan, M. P.; Culbert, A. A.; Cross, D. A.; Corcoran, S. L.; Yates, J. W.; Pearce, N. J.; Rausch, O. L.; Murphy, G. J.; Carter, P. S.; Cox, L. R.; Mills, D.; Brown, M. J.; Haigh, D.; Ward, R. W.; Smith, D. G.; Murray, K. J.; Reigh, A. D.; Holder, J. C. Selective small molecule inhibitors of glycogen synthase kinase-3 modulate glycogen metabolism and gene transcription. Chem. Biol. 2000, 7, 793-803.

27. Cline, G. W.; Johnson, K.; Regittnig, W.; Perret, P.; Tozzo, E.; Xiao, L.; Damico, C.; Shulman, G. I. Effects of a novel glycogen synthase kinase-3 inhibitor on insulin-stimulated glucose metabolism in zucker diabetic fatty rats. Diabetes 2002, 51, 2903-2910.

28. Stukenbrock, H.; Mussmann, R.; Geese, M.; Ferandin, Y.; Lozach, O.; Lemcke, T.; Kegel, S.; Lomow, A.; Burk, U.; Dohrmann, C.; Meijer, L.; Austen, M.; Kunick, C. 9-Cyano-1-azapaullone (cazpaullone), a glycogen synthase kinase-3 (GSK-3) inhibitor activating pancreatic beta cell protection and replication. J. Med. Chem. 2008, 51, 2196-2207.

29. Meijer, L.; Skaltsounis, A.-L.; Magiatis, P.; Polychronopoulos, P.; Knockaert, M.; Leost, M.; Ryan, X. P.; Vonica, C. A.; Brivanlou, A.; Dajani, R.; Crovace, C.; Tarricone, C.; Musacchio, A.; Roe, S. M.; Pearl, L.; Greengard, P. GSK-3-selective inhibitors derived from tyrian pruple indirubins. Chem. Biol. 2003, 10, 1255-1266.

30. Meijer, L.; Thunnissen, A.-M.; White, A. W.; Garnier, M.; Nikolic, M.; Tsai, L. H.; Walter, J.; Cleverley, K. E.; Salinas, P. C.; Wu, Y. Z.; Beirnat, J.; Mandelkow, E. M.; Kim, S. H.; Pettit, G. R. Inhibition of cyclin-dependent kinases, GSK-3beta and CK1 by hymenialdisine, a marine sponge constituent. Chem. Biol. 2000, 7, 51-63.

31. Bhat, R.; Xue, Y.; Berg, S.; Hellberg, S.; Ormo, M.; Nilsson, Y.; Radesater, A.-C.; Jerning, E.; Markgren, P.-O.; Borgegard, T.; Nylof, M.; Gimenez-Cassina, A.; Hernandez, F.; Lucas, J. J.; Diaz-Nido, J.; Avila, J. Structural insights and biological effects of glycogen synthase kinase 3-specific inhibitor AR-A014418. J. Biol. Chem. 2003, 278, 45937-45945.

32. Gragg, G. M.; Grothaus, P. G.; Newman, D. J. Impact of natural products on developing new anti-cancer agents. Chem. Rev. 2009, 109, 3012-3043.

33. Mazanetz, M. P.; Fischer, P. M. Untangling tau hyperphosphorylation in drug design for neurodegenerative disease. Nat. Rev. Drug Discov. 2007, 6, 464-479.

34. Martinez, A.; Alonso, M.; Castro, A.; Prez, C.; Moreno, F. J. First non-ATP competitive glycogen synthase kinase 3 beta inhibitors: thiadiazolidinones (TDZD) as potential drugs for the treatment of Alzheimer's disease. J. Med. Chem. 2002, 45, 1292-1299.

35. Martinez, A.; Alonso, M.; Castro, A.; Dorronsoro, I.; Gelpi, J. L.; Luque, J.; Perez, C.; Moreno, F. J. SAR and 3D-QSAR studies on thiadiazolidinone derivatives: exploration of structural requirements for glycogen synthase kinase 3 inhibitors. J. Med. Chem. 2005, 48, 7103-7112.

36. Kaidanovich-Beilin, O.; Milman, A.; Weizman, A.; Pick, C. G.; Eldar-Finkelman, H. Rapid antidepressive-like activity of specific glycogen synthase kinase-3 inhibitor and its effect on beta-catenin in mouse hippocampus. Biol. Psychiatry 2004, 55, 781-784.

37. Plotkin, B.; Kaidanovich, O.; Talior, I.; Eldar-Finkelman, H. Insulin mimetic action of synthetic phosphorylated peptide inhibitors of glycogen synthase kinase-3. J. Pharmacol. Exp. Ther. 2003, 305, 974-980.

38. Kaidanovich, O.; Eldar-Finkelman, H. Long-term treatment with novel glycogen synthase kinase-3 inhibitor improves glycose homeostatis in ob/ob mice: molecular characterization in liver and muscle. J. Pharmacol. Exp. Ther. 2006, 316, 17-24.

39. Berman, H. M. The protein data bank: a historical perspective. Acta Cryst. 2008, 64, 88-95.

40. Aoki, M.; Yokota, T.; Sugiura, I.; Sasaki, C.; Hasegawa, T.; Okumura, C.; Ishiguro, K.; Kohno, T.; Sugio, S.; Matsuzaki, T. Structural insight into nucleotide recognition in tau-protein kinase I/glycogen synthase kinase 3. Acta Crystallogr. 2004, 60, 439-446.

41. Bertrand, J. A.; Thieffine, S.; Vulpetti, A.; Cristinani, C.; Valsasina, B.; Knapp, S.; Kalisz, H. M.; Flocco, M. Structural characterization of the GSK-3beta active site using selective and non-selective ATP-minetic inhibitors. J. Mol. Biol. 2003, 333, 393-407.

42. Bhat, R.; Xue, Y.; Berg, S.; Hellberg, S.; Ormo, M.; Nilsson, Y.; Radesater, A.; Jerning, E.; Markgren, P.; Borgegard, T.; Nylof, M.; Gimenez-Cassina, A.; Hernandez, F.; Lucas, J. J.;

Diaz-Nido, J.; Avila, J. Structural insights and biological effects of glycogen synthase kinase 3-specific inhibitor AR-A014418. J. Biol. Chem. 2003, 278, 45937-45945.

43. Allard, J.; Nikolcheva, T.; Gong, L.; Wang, J.; Dunten, P.; Avnur, Z.; Waters, R.; Sun, Q.; Skinner, B. From genetics to therapeutics: the Wnt pathway and osteoporosis. To be Published.

44. Meijer, L.; Skaltsounis, A. L.; Magiatis, P.; Polychronopoulos, P.; Knockaert, M.; Leost, M.; Ryan, X. P.; Vonica, C. A.; Brivanlou, A.; Dajani, R.; Crovace, C.; Tarricone, C.; Musacchio, A.; Roe, S. M.; Pearl, L.; Greengard, P. GSK-3-selective inhibitors derived from tyrian purple indirubins. Chem. Biol. 2003, 10, 1255-1266.

45. Shin, D.; Lee, S. C.; Heo, Y. S.; Lee, W. Y.; Cho, Y. S.; Kim, Y. E.; Hyun, Y. L.; Cho, J. M.; Lee, Y. S.; Ro, S. Design and synthesis of 7-hydroxy-1H-benzoimidazole derivatives as novel inhibitors of glycogen synthase kinase-3beta. Bioorg. Med. Chem. Lett. 2007, 17, 5686-5689.

46. Zhang, H. C.; Bonaga, L. V.; Ye, H.; Derian, C. K.; Damiano, B. P.; Maryanoff, B. E. Novel bis(indolyl)maleimide pyridinophanes that are potent, selective inhibitors of glycogen synthase kinase-3. Bioorg. Med. Chem. Lett. 2007, 17, 2863-2868.

47. Atilla-Gokcumen, G. E.; Pagano, N.; Streu, C.; Maksimoska, J.; Filippakopoulos, P.; Knapp, S.; Meggers, E. Extremely tight binding of a ruthenium complex to glycogen synthase kiinase 3. ChemBioChem 2008, 9, 2933-2936.

48. Menichincheri, M.; Bargiotti, A.; Berthelsen, J.; Bertrand, J. A.; Bossi, R.; Ciavolella, A.; Cirla, A.; Cristiani, C.; Croci, V.; D'Alessio, R.; Fasolini, M.; Fiorentini, F.; Forte, B.; Isacchi, A.; Martina, K.; Molinari, A.; Montagnoli, A.; Orsini, P.; Orzi, F.; Pesenti, E.; Pezzetta, D.; Pillan, A.; Poggesi, I.; Roletto, F.; Scolaro, A.; Tato, M.; Tibolla, M.; Valsasina, B.; Varasi, M.;

Volpi, D.; Santocanale, C.; Vanotti, E. First Cdc7 kinase inhibitors: pyrrolopyridinones as potent and orally active antitumor agents. 2. Lead discovery. J. Med. Chem. 2009, 52, 293-307.

49. Saitoh, M.; Kunitomo, J.; Kimura, E.; Hayase, Y.; Kobayashi, H.; Uchiyama, N.; Kawamoto, T.; Tanaka, T.; Mol, C. D.; Dougan, D. R.; Textor, G. S.; Snell, G. P.; Itoh, F. Design, synthesis and structure-activity relationships of 1,3,4-oxadiazole derivatives as novel inhibitors of glycogen synthase kinase-3beta. Bioorgan. Med. Chem. 2009, 17, 2017-2029.

50. Saitoh, M.; Kunitomo, J.; Kimura, E.; Iwashita, H.; Uno, Y.; Onishi, T.; Uchiyama, N.; Kawamoto, T.; Tanaka, T.; Mol, C. D.; Dougan, D. R.; Textor, G. P.; Snell, G. P.; Takizawa, M.; Itoh, F.; Kori, M. 2-{3-[4-(Alkylsulfinyl)phenyl]-1-benzofuran-5-yl}-5-methyl-1,3,4-oxadiazole derivatives as novel inhibitors of glycogen synthase kinase-3beta with good brain permeability. J. Med. Chem. 2009, 52, 6270-6286.

51. Aronov, A. M.; Tang, Q.; Martinez-Botella, G.; Bermis, G. W.; Cao, J.; Chen, G.; Ewing, N. P.; Ford, P. J.; Germann, U. A.; Green, J.; Hale, M. R.; Jacobs, M.; Janetka, J. W.; Maltais, F.; Markland, W.; Namchuk, M. N.; Nanthakumar, S.; Poondru, S.; Straub, J.; Ter Haar, E.; Xie, X. Structure-guided design of potent and selective pyrimidylpyrrole inhibitors of extracellular signal-regulated kinase (ERK) using conformational control. J. Med. Chem. 2009, 52, 6362-6368.

52. Arnost, M.; Pierce, A.; Haar, E. T.; Lauffer, D.; Madden, J.; Tanner, K.; Green, J. 3-Aryl-4-(arylhydrazono)-1H-pyrazol-5-ones: highly ligand efficient and potent inhibitors of GSK3beta. Bioorg. Med. Chem. Lett. 2010, 20, 1661-1664.

53. Atilla-Gokcumen, G. E.; Di Costanzo, L.; Meggers, E. Structure of anticancer ruthenium half-sandwich complex bound to glycogen synthase kinase 3. J. Biol. Inorg. Chem. 2010, 16, 45-50.

54. Feng, L.; Geisselbrecht, Y.; Blanck, S.; Wilbuer, A.; Atilla-Gokcumen, G. E.; Filippakopoulos, P.; Kraling, K.; Celik, M. A.; Harms, K.; Maksimoska, J.; Marmorstein, R.; Frenking, G.; Knapp, S.; Essen, L. O.; Meggers, E. Structurally sophisticated octahedral metal complexes as highly selective protein kinase inhibitors. J. Am. Chem. Soc. 2011, 133, 5976-5986.

55. Kim, H.-J.; Choo, H.; Cho, Y. S.; No, K. T.; Pae, A. N. Novel GSK-3beta inhibitors from sequential virtual screening. Bioorgan. Med. Chem. 2008, 16, 636-643.

56. Dessalew, N.; Bharatam, P. V. Investigation of potential glycogen synthase kinase 3 inhibitors using pharmacophore mapping and virtual screening. Chem. Biol. Drug Des. 2006, 68, 154-165.

57. Taha, M. O.; Bustanji, Y.; Al-Ghussein, M. A. S.; Mohammad, M.; Zalloum, H. Pharmacophore modeling, quantitative structure-activity relationship analysis, and in silico screening reveal potent glycogen synthase kianse-3beta inhibitory activities for cimetidine, hydroxychloroquine, and gemifloxacin. J. Med. Chem. 2008, 51, 2062-2077.

58. Gadakar, P. K.; Phukan, S.; Dattatreya, P.; Balaji, V. N. Pose prediction accuracy in docking studies and enrichment of actives in the active site of GSK-3beta. J. Chem. Inf. Model. 2007, 47, 1446-1459.

59. Lesuisse, D.; Dutruc-Rosset, G.; Tiraboschi, G.; Dreyer, M. K.; Maignan, S.; Chevalier, A.; Halley, F.; Bertrand, P.; Burgevin, M.-C.; Quarteronet, D.; Rooney, T. Rational design of potent GSK3beta inhibitors with selectivity for Cdk1 and Cdk2. Bioorgan. Med. Chem. 2010, 20, 1985-1989.

60. Patel, D. S.; Dessalew, N.; Iqbal, P.; Bharatam, P. V. Structure-based approaches in the design of GSK-3 selective inhibitors. Curr. Protein Pept.Sc. 2007, 8, 352-364.

61. Smith, D. G.; Buffet, M.; Fenwick, A. E.; Haigh, D.; Ife, R. J.; Saunders, M.; Slingsby, B. P.; Stacey, R.; Ward, R. W. 3-Anilino-4-arylmaleimides: potent and selective inhibitors of glycogen synthase kinase-3 (GSK-3). Bioorg. Med. Chem. Lett. 2001, 11, 635-639.

62. Leclerc, S.; Garnier, M.; Hoessel, R.; Marko, D.; Bibb, J. A.; Snyder, G. L.; Greengard, P.; Biernat, J.; Wu, Y. Z.; Mandelkow, E. M.; Eisenbrand, G.; Meijer, L. Indirubins inhibit glycogen synthase kinase-3beta and CDK5/P25, two protein kinases involved in abnormal tau phosphorylation in Alzheimer's disease. J. Biol. Chem. 2001, 276, 251-260.

63. Polychronopoulos, P.; Magiatis, P.; Skaltsounis, A. L.; Myrianthopoulos, V.; Mikros, E.; Tarricone, A.; Musacchio, A.; Roe, S. M.; Pearl, L.; Leost, M.; Greengard, P.; Meijer, L. Structural basis of the synthesis of indirubins as potent and selective inhibitors of glycogen synthase kinase-3 and cyclin-dependent kinases. J. Med. Chem. 2004, 47, 935-946.

64. Kunich, C.; Lauenroth, K.; Wieking, K.; Xie, X.; Schultz, C.; Gussio, R.; Zaharevitz, D.; Leost, M.; Meijer, L.; Weber, A.; Jorgensen, F. S.; Lemcke, T. Evaluation and comparison of 3D-QSAR CoMSIA models for CDK1, CDK5, and GSK-3 inhibition by paullones. J. Med. Chem. 2004, 47.

65. Mettey, Y.; Gompel, M.; Thomas, V.; Garnier, M.; Leost, M.; Ceballos-Picot, I.; Noble, M.; Endicott, J.; Vierfond, J. M.; Meijer, L. Aloisines, a new family of CDK/GSK-3 inhibitors. SAR study, crystal structure in complex with CDK2, enzyme selectivity, and cellular effects. J. Med. Chem. 2003, 46, 222-236.

66. Peat, A. J.; Boucheron, J. A.; Dickerson, S. H.; Garrido, D.; Mills, W.; Peckham, J.; Preugschat, F.; Smalley, T.; Schweiker, S. L.; Wilson, J. R.; Wang, T. Y.; Zhou, H. Q.; Thomas, S. A. Novel pyrazolopyrimidine derivatives as GSK-3 inhibitors. Bioorg. Med. Chem. Lett. 2004, 14, 2121-2125.

67. Peat, A. J.; Garrido, D.; Boucheron, J. A.; Schweiker, S. L.; Dickerson, S. H.; Wilson, J. R.; Wang, T. Y.; Thomas, S. A. Novel GSK-3 inhibitors with improved cellular activity. Bioorg. Med. Chem. Lett. 2004, 14, 2127-2130.

68. Witherington, J.; Bordas, V.; Haigh, D.; Hickey, D. M. B.; Ife, R. J.; Rawlings, A. D.; Slingsby, B. P.; Smith, D. G.; Ward, R. W. 5-Aryl-pyrazolo[3,4-b]pyridazines: potent inhibitors of glycogen synthase kinase-3 (GSK-3). Bioorg. Med. Chem. Lett. 2003, 13, 1581-1584.

69. Tavares, F. X.; Boucheron, J. A.; Dickerson, S. H.; Griffin, R. J.; Preugschat, F.; Thomson, S. A.; Wang, T. Y.; Zhou, H. Q. N-phenyl-4-pyrazolo[1,5-b]pyridazin-3-ylpyrimidin-2-amines as potent and selective inhibitors of glycogen synthase kinase 3 with good cellular efficacy. J. Med. Chem. 2004, 47, 4716-4730.

70. Witherington, J.; Bordas, V.; Gaiba, A.; Garton, N. S.; Naylor, A.; Rawlings, A. D.; Slingsby, B. P.; Smith, D. G.; Takle, A. K.; Ward, R. W. 6-Aryl-pyrazolo[3,4-b]pyridines: potent inhibitors of glycogen synthase kinase-3 (GSK-3). Bioorg. Med. Chem. Lett. 2003, 13, 3055-3057.

71. Witherington, J.; Bordas, V.; Gaiba, A.; Naylor, A.; Rawlings, A. D.; Slingsby, B. P.; Smith, D. G.; Takle, A. K.; Ward, R. W. 6-Heteroaryl-pyrazolo[3,4-b]pyridines: potent and selective inhibitors of glycogen synthase kinase-3 (GSK-3). Bioorg. Med. Chem. Lett. 2003, 13, 3059-3062.

72. Witherington, J.; Bordas, V.; Garland, S. L.; Hickey, D. M. B.; Ife, R. J.; Liddle, J.; Saunders, M.; Smith, D. G.; Ward, R. W. 5-Aryl-pyrazolo[3,4-b]pyridines: potent inhibitors of glycogen synthase kinase-3 (GSK-3). Bioorg. Med. Chem. Lett. 2003, 13, 1577-1580.

73. Gaisina, I. N.; Gallier, F.; Ougolkov, A. V.; Kim, K. H.; Kurome, T.; Guo, S.; Holzle, D.; Luchini, D. N.; Blond, S. Y.; Billadeau, D. D.; Kozikowski, A. P. From a natural product lead to

the identification of potent and selective benzofuran-3-yl-(indol-3-yl)maleimides as glycogen synthase kinase 3beta inhibitors that suppress proliferation and survival of pancreatic cancer cells. J. Med. Chem. 2009, 52, 1853-1863.

74. Katritzky, A. R.; Pacureanu, L. M.; Dobchev, D. A.; Fara, D. C.; Duchowicz, P. R.; Karelson, M. QSAR modeling of the inhibition of glycogen synthase kinase-3. Bioorgan. Med. Chem. 2006, 14, 4987-5002.

75. Sivaprakasam, P.; Xie, A.; Doerksen, R. J. Probing the physicochemical and structural requirements for glycogen synthase kinase-3alpha inhibition: 2D-QSAR for 3-anilino-4-phenylmaleimides. Bioorg. Med. Chem. 2006, 14, 8210-8218.

76. Sivaprakasam, P.; Daga, P. R.; Xie, A.; Doerksen, R. J. Glycogen synthase kinase-3 inhibition by 3-anilino-4-phenylmaleimides: insights from 3D-QSAR and docking. J. Comput. Aided Mol. Des. 2009, 23, 113-127.

77. Zeng, M.; Jiang, Y.; Zhang, B.; Zheng, K.; Zhang, N.; Yu, Q. 3D QSAR studies on GSK-3 inhibition by aloisines. Bioorg. Med. Chem. Lett. 2005, 15, 395-399.

78. Zhang, N.; Jiang, Y.; Zou, J.; Zhang, B.; Jin, H.; Wang, Y.; Yu, Q. 3D QSAR for GSK-3beta inhibition by indirubin analogues. Eur. J. Med. Chem. 2006, 41, 373-378.

79. Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine learning in virtual screening. Comb. Chem. High T. Scr. 2009, 12, 332-343.

80. Taha, M.; Bustanji, Y.; Al-Ghussein, M. A. S.; Mohammad, M.; Zalloum, H.; Al-Masri, I. M.; Atallah, N. Pharmacophore modeling, quantitative structure-activity relationship analysis, and in silico screening reveal potent glycogen synthase kinase-3beta inhibitory activities for cimetidine, hydroxychloroquine, and gemifloxacin. J. Med. Chem. 2008, 51, 2062-2077.

81. Goodarzi, M.; Freitas, M. P.; Jensen, R. Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3beta inhibitory activities. J. Chem. Inf. Model. 2009, 49, 824-832.

82. Motta, C. L.; Sartini, S.; Tuccinardi, T.; Nerini, E.; Settimo, F. D.; Martinelli, A. Computational studies of epidermal growth factor receptor: docking reliability, three-dimensional quantitative structure-activity relationship analysis, and virtual screening studies. J. Med. Chem. 2009, 52, 964-975.

83. Fang, J.; Huang, D.; Zhao, W.; Ge, H.; Luo, H. B.; Xu, J. A new protocol for predicting novel GSK-3beta ATP competitive inhibitors. J. Chem. Inf. Model. 2011, 51, 1431–1438.

84. Patel, D. S.; Bharatam, P. V. New leads for selective GSK-3 inhibition: pharmacophore mapping and virtual screening studies. J. Comput. Aided Mol. Des. 2006, 20, 55-66.

85. Barril, X.; Hubbard, R. E.; Morley, S. D. Virtual screening in structure-based drug design. Mini-Rev. Med. Chem. 2004, 4, 779-791.

86. Kang, N. S.; Lee, G. N.; Kim, C. H.; Bae, M. A.; Kim, I.; Cho, Y. S. Identification of small molecules that inhibit GSK-3beta through virtual screening. Bioorg. Med. Chem. Lett. 2009, 19, 533-537.

87. Polgar, T.; Baki, A.; Szendrei, G. I.; Keseru, G. M. Comparative vritual and experiemental high-throughput screening for glycogen synthase kinase-3beta inhibitors. J. Med. Chem. 2005, 48, 7946-7959.

88. Dessalew, N.; Bharatam, P. V. Structure based de novo design of novel glycogen synthase kinase 3 inhibitors. Bioorg. Med. Chem. 2007, 15, 3728-3736.

89. Roberts, B. C.; Mancera, R. L. Ligand-protein docking with water molecules. J. Chem. Inf. Model. 2008, 48, 397-408.

90. Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. Proc. Natl. Acad. Sci. U.S.A. 2007, 104, 808-813.

91. Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the active-site solvent in the thermodynamics of Factor Xa Ligand Binding. J. Am. Chem. Soc. 2008, 130, 2817-2831.

92. Robinson, D. D.; Sherman, W.; Farid, R. Understanding kinase selectivity through energetic analysis of binding site waters. CHemMEDCHEM. 2010, 5, 618-627.

93. Shaltiel, S.; Cox, S.; Taylor, S. S. Conserved water molecules contribute to the extensive network of interactions at the active site of protein kinase A. Proc. Natl. Acad. Sci. U.S.A. 1997, 95, 484-491.

94. Knight, J. D. R.; Hamelberg, D.; Andrew, M. J.; Kothary, R. The role of conserved water molecules in the catalytic domain of protein kinases. Proteins 2009, 76, 527-535.

95. Lu, S.; Jiang, y.; Lv, j.; Zou, j.; Wu, T. Role of bridging water molecules in GSK3beta-inhibitor complexes: insight from QM/MM, MD, and molecular docking studies. J. Comput. Chem. 2011, 32, 1907-1918.

96. Vulpetti, A.; Crivori, P.; Cameron, A.; Bertrand, J.; Brasca, M. G.; D'Alessio, R.; Pevarello, P. Structure-based approaches to improve selectivity: CDK2-GSK3beta binding site analysis. J. Chem. Inf. Model. 2005, 45, 1282-1290.

97. Lesuisse, D.; Dutruc-Rosset, G.; Tiraboschi, G.; Dreyer, M.; Maignan, S.; Chevalier, A.; Halley, F.; Bertrand, P.; Burgevin, M. C.; Quarteronet, D.; Rooney, T. Rational design of potent GSK3beta inhibitors with selectivity for cdk1 and cdk2. Bioorg. Med. Chem. Lett. 2010, 20, 1985-1989.

98. Chen, Q.; Cui, W.; Cheng, Y.; Zhang, F.; Ji, M. Studying the mechanism that enables paullones to selectively inhibit glycogen synthase kinase 3 rather than cyclin-dependent kinase 5 by molecular dynamics simulations and free-energy calculations. Journal of Molecular Modeling 2011, 1-9.

99. Ilouz, R.; Kowalsman, N.; Eisenstein, M.; Eldar-Finkelman, H. Identification of novel glycogen syntahse kinase-3beta substrate-interacting residues suggests a common mechanism for substrate recognition. J. Biol. Chem. 2006, 281, 30621-30630.

100. Zhang, N.; Jiang, y.; Zou, j.; Yu, Q.; Zhao, W. Structural basis for the complete loss of GSK3beta catalytic activity due to R96 mutation investigated by molecular dynamics study. Proteins 2009, 75, 671-681.

101. Buch, I.; Fishelovitch, D.; London, N.; Raveh, B.; Wlfson, H. J.; Nussinov, R. Allosteric regulation of glycogen synthase kinase 3beta: a theoretical study. Biochemistry 2010, 49, 10890-10901.

102. Lu, S.; Jiang, Y.; Zou, j.; Wu, T. Molecular modeling and molecular dynamics simulation studies of the GSK3beta/ATP/substrate complex: understanding the unique P+4 primed phosphorylation specificity for GSK3beta substrates. J. Chem. Inf. Model. 2011, 51, 1025-1036.

103. Zhang, N.; Jiang, y.; Zou, J.; Zhuang, S.; Jin, H.; Yu, Q. Insights into unbinding mechanisms upon two mutations investigated by molecular dynamics study of GSK3beta-axin complex: role of packing hydorphobic residues. Proteins 2007, 67, 941-949.

104. Tang, X.; Lo, C.; Chuang, Y.; Chen, C.; Sun, Y.; Hong, Y.; Yang, C. Prediction of the binding mode between GSK3beta and a peptide derived from GSKIP using molecular dynamics simulation. Biopolymers 2011, 95, 461-471.

105. Phase, version 3.2; Schrődinger, LLC: New York, NY, 2010.

106. MacroModel, version 9.8; Schrődinger, LLC: New York, NY, 2010.

107. Lum, C.; Kahl, J.; Kessler, L.; Kucharski, J.; Lundstrom, J.; Miller, S.; Nakanishi, H.; Pei, Y.; Pryor, K.; Roberts, E.; Sebo, L.; Sullivan, R.; Urban, J.; Wang, Z. 2,5-Diaminopyrimidines and 3,5-disubstituted azapurines as inhibitors of glycogen synthase kinase-3 (GSK-3). Bioorg. Med. Chem. Lett. 2008, 18, 3578-3581.

108. Ha, H.-H.; Kim, J. S.; Kim, B. M. Novel heterocycle-substituted pyrimidines as inhibitors of NF-kappaB transcription regulation related to TNF-alpha cytokine release. Bioorg. Med. Chem. Lett. 2008, 18, 653-656.

109. Peat, A. J.; Boucheron, J. A.; Dickerson, S. H.; Garrido, D.; Mills, W.; Peckham, J.; Preugschat, F.; Smalley, T.; Schweiker, S. L.; Wilson, J. R.; Wang, T. Y.; Zhou, H. Q.; Thomson, S. A. Novel pyrazolopyrimidine derivatives as GSK-3 inhibitors. Bioorg. Med. Chem. Lett. 2004, 14, 2121-2125.

110. Engler, T. A.; Henry, J. R.; Malhotra, S.; Cunningham, B.; Furness, K.; Brozinick, J.; Burkholder, T. P.; Clay, M. P.; Clayton, J.; Diefenbacher, C.; Hawkins, E.; Iversen, P. W.; Li, Y.; Lindstrom, T. D.; Marquart, A. L.; McLean, J.; Mendel, D.; Misener, E.; Briere, D.; O'Toole, J. C.; Porter, W. J.; Queener, S.; Reel, J. K.; Owens, R. A.; Brier, R. A.; Eessalu, T. E.; Wagner, J. R.; Campbell, R. M.; Vaughn, R. Substituted 3-imidazol[1,2-a]pyridin-3-yl-4-(1,2,3,4-tetrahydro-[1,4]diazepino-[6,7,1-hi]indol-7-yl)pyrrole-2,5-diones as highly selective and potent inhibitors of glycogen synthase kinase-3. J. Med. Chem. 2004, 47, 3934-3937.

111. Olesen, P. H.; Sorensen, A. R.; Urso, B.; Kurtzhals, P.; Bowler, A. N.; Ehrbar, U.; Hansen, B. F. Synthesis and in vitro characterization of 1-(4-aminofurazan-3-yl)-5-dialkylaminomethyl-1H-[1,2,3]triazole-4-carboxylic acid derivatives. A new class of selective GSK-3 inhibitors. J. Med. Chem. 2003, 26, 3333-3341.

112. Maeda, Y.; Nakano, M.; Sato, H.; Miyazaki, Y.; Schweiker, S. L.; Smith, J. L.; Truesdale, A. T. 4-Acylamino-6-arylfuro[2,3-d]pyrimidines: potent and selective glycogen synthase kinase-3 inhibitors. Bioorg. Med. Chem. Lett. 2004, 14, 3907-3911.

113. Ring, D. B.; Johnson, K. W.; Henriksen, E. J.; Nuss, J. M.; Goff, D.; Kinnick, T. R.; Ma, S. T.; Reeder, J. W.; Samuels, I.; Slabiak, T.; Wagman, A. S.; Hammond, M.-E. W.; Harrison, S. D. Selective glycogen synthase kinase 3 inhibitors potentiate insulin activation of glucose transport and utilization in itro and in vivo. Diabetes 2003, 52, 588-595.

114. Kozikowski, A. P.; Gaisina, I. N.; Yuan, H.; Petukhov, P. A.; Blond, S. Y.; Fedolak, A.; Caldarone, B.; McGonigle, P. Structure-based design leads to the identification of lithium mimetics that block mania-like effects in rodents. Possible new GSK-3beta therapies for bipolar disorders. J. Am. Chem. Soc. 2009, 129, 8328-8332.

115. Zhang, H.-C.; Ye, H.; Conway, B. R.; Derian, C. K.; Addo, M. F.; Kuo, G.-H.; Hecker, L. R.; Croll, D. R.; Li, J.; Westover, L.; Xu, J. Z.; Look, R.; Demarest, K. T.; Andrade-Gordon, P.; Damiano, B. P.; Maryanoff, B. E. 3-(7-Azaindolyl)-4-arylmaleimides as potent, selective inhibitors of glycogen synthase kinase-3. Bioorg. Med. Chem. Lett. 2004, 14, 3245-3250.

116. Haigh, J. A.; Pickup, B. T.; Crant, J. A.; Nicholls, A. Small molecule shape-fingerprints. J. chem. Inf. Model. 2005, 45, 673-684.

117. Leach, A. R.; Gillet, V. J. An introduction to chemoinformatics. Kluwer academic Publishers: Dordrecht, the Netherlands, 2003.

118. Canvas, version 1.3; Schrődinger, LLC: New York, NY, 2010.

119. Glide, version 5.6; Schrődinger, LLC: New York, NY, 2010.

120. Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. 1995, 245, 43-53.

121. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 1997, 267, 727-748.

122. Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein-ligand docking using GOLD. J. Med. Chem. 2005, 48, 6504-6515.

123. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. J. Comput. Chem. 2007, 28, 1145-1152.

124. Morris, G. M.; Goodsell, D. S.; Haliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a lamerckian genetic algorithm and empirical binding free energy function. J. Comput. Chem. 1998, 19, 1639-1662.

125. Molecular Operating Environment, MOE 2010.10; Chemical Computing Group, Inc: Montreal, Quebec, Canada, 2010.

126. Atilla-Gokcumen, G. E.; Pagano, N.; Streu, C.; Maksimoska, J.; Filippakopoulos, P.; Knapp, S.; Meggers, E. Extremely tight binding of a ruthenium complex to glycogen synthase kiinase 3. To be Published.

127. Saitoh, M.; Kunitomo, J.; Kimura, E.; Hayase, Y.; Kobayashi, H.; Uchiyama, N.; Kawamoto, T.; Tonaka, T.; Mol, C. D.; Dougan, D. R.; Textor, G. S.; Snell, G. P.; Itoh, F. Design, synthesis and structure-activity relationships of 1,3,4-oxadiazole derivatives as novel inhibitors of glycogen synthase kinase-3beta. Bioorgan. Med. Chem. 2009, 17, 2017-2029.

128. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J. Comp. Chem. 1996, 17, 490-519.

129. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M.; Perry, J. K.; Francis, P.; Shenkin, P. S.

Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 2004, 47, 1739-1749.

130. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J. Med. Chem. 2004, 47, 1750-1759.

131. Kuo, G.-H.; Prouty, C.; DeAngelis, A.; Shen, L.; O'Neill, D. J.; Shah, C.; Connolly, P. J.; Murray, W. V.; Conway, B. R.; Cheung, P.; Westover, L.; Xu, J. Z.; Look, R.; Demarest, K. T.; Emanuel, S.; Middleton, S. A.; Jolliffe, L.; Beavers, M. P.; Chen, X. Synthesis and discovery of macrocyclic polyoxygenated bis-7-azaindolylmaleimides as a novel series of potent and highly selective glycogen sysnthase kinase-3beta inhibitors. J. Med. Chem. 2003, 46, 4021-4031.

132. LigPrep, version 2.4; Schrődinger, LLC: New York, NY, 2010.

133. Bain, J.; McLauchlan, H.; Elloitt, M.; Cohen, P. The specificities of protein kinase inhibitors: an update. Biochem. J. 2003, 371, 199-204.

134. Prasanna, S.; Daga, P. R.; Xie, A.; Doerksen, R. J. Glycogen synthase kinase-3 inhibition by 3-anilino-4-phenylmaleimides: insights from 3D-QSAR and docking. J. Comput. Aided Mol. Des. 2009, 23, 113-127.

135. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. J. Med. Chem. 2006, 49, 534-553.

136. Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ. Res. 2009, 20, 241-266.

137. Coghlan, M. P.; Culbert, A. A.; Cross, D. A.; Corcoran, S. L.; Yates, J. W.; Pearce, N. J.; Rausch, O. L.; Murphy, G. J.; Carter, P. S.; Cox, L. R.; Mills, D.; Brown, M. J.; Haigh, D.;

Ward, R. W.; Smith, D. G.; Murray, K. J.; Reith, A. D.; Holder, J. C. Selective small molecule inhibitors of glycogen synthase kinase-3 modulate glycogen metabolism and gene transcription. Chem. Biol. 2000, 7, 793-803.

138. Kozikowski, A. P.; Gaisina, I. N.; Petukhov, P. A.; Sridhar, J.; King, L. T.; Blond, S. Y.; Duka, T.; Rusnak, M.; Sidhu, A. Highly potent and specific GSK-3beta inhibitors that block tau phosphorylation and decrease alpha-synuclein protein expression in a cellular model of Parkinson's disease. CHEMMEDCHEM. 2006, 1, 256-266.

139. Engler, T. A.; Henry, J. R.; Malhotra, S.; Cunningham, B.; Furness, K.; Brozinick, J.; Burkholder, T. P.; Clay, M. P.; Clayton, J.; Diefenbacher, C.; Hawkins, E.; Iversen, P. W.; Li, Y.; Lindstrom, T. D.; Marquart, A. L.; McLean, J.; Mendel, D.; Misener, E.; Briere, D.; O'Toole, J. C.; Porter, W. J.; Queener, S.; Reel, J. K.; Owens, R. A.; Brier, R. A.; Eiessalu, T. E.; Wagner, J. R.; Campbell, R. M.; Vaughn, R. Substituted 3-imidazo[1,2-a]pyridin-3-yl-4-(1,2,3,4-tetrahydro-[1,4]diazepino[6,7,1-hi]indol-7-yl)pyrrole-2,5-diones as highly selective and potent inhibitors of glycogen synthase kinase-3. J. Med. Chem. 2004, 47, 3934-3937.

140. Bain, J.; McLauchlan, H.; Elliott, M.; Cohen, P. The specificities of protein kinase inhibitors: an update. Biochem. J. 2003, 371, 199-204.

141. Leost, M.; Schultz, C.; Link, A.; Wu, Y. Z.; Biernat, J.; Mandelkow, E. M.; Bibb, J. A.; Snyder, G. L.; Greengard, P.; Zaharevitz, D.; Gussio, R.; Senderowicz, A. M.; Sausville, E. A.; Kunick, C.; Meijer, L. Paullones are potent inhibitors of glycogen synthase kinase-3beta and cyclin-dependent kinase 5/p25. Eur. J. Biochem. 2000, 267, 5983-5994.

142. Vougogiannopoulou, K.; Ferandin, Y.; Bettayeb, K.; Myrianthopoulos, V.; Lozach, O.; Fan, Y. Z.; Johnson, C. H.; Magiatis, P.; Skaltsounis, A. L.; Mikros, E.; Meijer, L. Soluble 3#,6-

substituted indirubins with enhanced selectivity toward glycogen synthase kinase-3 alter circadian period. J. Med. Chem. 2008, 51, 6421-6431.

143. Stukenbrock, H.; Mussmann, R.; Geese, M.; Ferandin, Y.; Lozach, O.; Lemcke, T.; Kegel, S.; Lomow, A.; Burk, U.; Dohrmann, C.; Meijer, L.; Austen, M.; C., K. 9-Cyano-1-azapaullone(cazpaullone), a glycogen synthase kinase-3(GSK-3) inhibitor activating pancreatic # cell protection and replication. J. Med. Chem. 2008, 51, 2196-2207.

144. Kozikowski, A. P.; Gaisina, I. N.; Yuan, H. B.; Petukhov, P. A.; Blond, S. Y.; Fedolak, A.; Caldarone, B.; McGonigle, P. Structure-based design leads to the identification of lithium mimetics that block mania-like effects in rodents. Possible new GSK-3beta therapies for bipolar disorders. J. Am. Chem. Soc. 2007, 129, 8328-8332.

145. Fu, G.; Sivaprakasam, P.; Dale, O. R.; Manly, S. P.; Cutler, S. J.; Doerksen, R. J. Pharmacophore modeling, ensemble docking, and virtual screening studies on glycogen synthase kinase-3beta. XXXX XXXX, XX, XXXX.

146. MOE, Chemical Computing Group: Montreal,Quebec, Canada, 2009.

147. DRAGON for linux (Software for Molecular Descriptor Calculations), Version 1.4; Talete srl: Milano, Italy, 2010.

148. Forgy, E. W. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 1965, 21, 768-769.

149. Golbraikh, A.; Tropsha, A. Beware of q2. J. Mol. Graph. Model. 2002, 20, 269-276.

150. Vapnik, V. The nature of statistical learning theory. Springer: New York, 1995.

151. Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273-297.

152. Drucker, H.; Burges, C.; Kaufman, L.; Smola, A.; Vapnik, V. In Support vector regression machines, Advances in Neural Information Processing Systems, 1997; Mozer, M.; Jordan, M.; Petsche, T., Eds. MIT Press: 1997; pp 155-161.

153. MATLAB, R2011a; The Mathworks Inc.: Natick, MA, 2011.

154. Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machines, 2011.

155. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007, 23, 2507-2517.

156. Kennedy, J.; Eberhart, R. C. In Particle swarm optimization, Proceedings of IEEE International Conference on Neural Networks, Perth, 1995; Perth, 1995; pp 1942-1948.

157. Kennedy, J.; Eberhart, R. C. In A new optimizer using particle swarm theory, Sixth International Symposim on Micro Machine and Human Science, Nagoya, 1995; Nagoya, 1995; pp 39-43.

158. Shi, Y.; Eberhart, R. C. In A modified particle swarm optimizer, Proceedings of IEEE International Conference on Evolutionary Computation, Anchorage, AK, USA, 1998; Anchorage, AK, USA, 1998; pp 69-73.

159. Eberhart, R. C.; Shi, Y. In Particle swarm optimization: developments, applications and resources, Proceedings of IEEE International Conference on Evolutionary Computation, Seoul, 2001; Seoul, 2001; pp 81-86.

160. Wang, X. Y.; Yang, J.; Teng, X. L.; Xia, W. J.; Jensen, R. Feature selection based on rough sets and particle swarm optimization. Pattern Recogn. Lett. 2007, 28, 459-471.

161. Agrafiotis, D. K.; Cedeno, W. Feature selection for structure-activity correlation using binary particle swarms. J. Med. Chem. 2002, 45, 1098-1107.

162. Huang, J.; Ma, G.; Muhammad, I.; Cheng, Y. Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. J. Chem. Inf. Model. 2007, 47, 1638-1647.

163. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5-32.

164. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 2003, 43, 1947-1958.

165. Jaiantilal, A. Classification and Regression by randomForest-matlab, 2009.

166. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set descriptor space. ATLA Altern. Lab. Anim. 2005, 33, 445-459.

167. Chen, Y.; Bi, J.; Wang, J. Z. MILES: Multiple-instance learning via embedded instance selection. IEEE T Pattern Anal 2006, 28, 1931-1947.

168. McGregor, M. J.; Muskal, M. M. Pharmacophore fingerprint. 1. Application to QSAR and focused library design. J. Chem. Inf. Comput. Sci. 1999, 39, 569-574.

169. Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. J. Med. Chem. 1999, 42, 3251-3264.

170. Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: description and application to alpha1-adrenergic antagonists. J. Med. Chem. 2000, 43, 2770-2774.

171. Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. J Med Chem 2002, 45, 1737-1740.

172. Li, W. X.; Li, L.; Eksterowicz, J.; Ling, X.; Cardozo, M. Significant analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. J. Chem. Inf. Model. 2007, 47, 2429-2438.

173. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. J Chem Inf Comp Sci 1998, 38, 983-996.

174. Canvas, version 1.4; Schrődinger, LLC: New York, NY, 2011.

175. Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. P Natl Acad Sci USA 2001, 98, 5116-5121.

176. Maron, O.; Lozano-Perez, T. A framework for multiple-instance learning. Adv Neur In 1998, 10, 570-576.

177. Bi, J.; Bennett, K. P.; Embrechts, M.; Breneman, C.; Song, M. Dimensionality reduction via sparse support vector machines. J Mach Learn Res 2003, 3, 1229-1243.

178. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975, 405, 442-451.

179. Green, J.; Marhefka, C. Phthalimide compounds useful as protein kinase inhibitors. WO 2005/039564 A1, 2005.

180. Lescot, E.; Bureau, R.; Santos, J. S. O.; Rochais, C.; Lisowski, V.; Lancelot, J.; Rault, S. 3D-QSAR and docking studies of selective GSK-3beta inhibitors. Comparison with a thieno[2,3-b]pyrrolizinone derivative, a new potential lead for GSK-3beta ligands. J. Chem. Inf. Model. 2005, 45, 708-715.

181. Dessalew, N.; Patel, D. S.; Bharatam, P. V. 3D-QSAR and molecular docking studies on pyrazolopyrimidine derivatives as glycogen synthase kinase-3beta inhibitors. J. Mol. Graph. Model. 2007, 25, 885-895.

182. Lather, V.; Kristam, R.; Saini, J. S.; Kristam, R.; Karthikeyan, N. A.; Balaji, V. N. QSAR Models for prediction of glycogen synthase kinase-3beta inhibitory activity of indirubin derivatives. QSAR Comb. Sci. 2008, 6, 718-728.

183. Prasanna, S.; Daga, P. R.; Xie, A.; Doerksen, R. J. Glycogen synthase kinase-3 inhibition by 3-anilino-4-phenylmaleimides: insights from 3D-QSAR and docking. J. Comput. Aided Mol. Des. 2009, 23, 113-127.

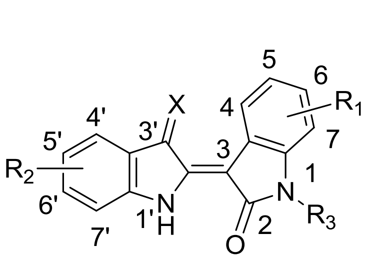
APPENDIX

Appendix: A. Summary of the previous QSAR studies related to GSK-3 inhibitors.

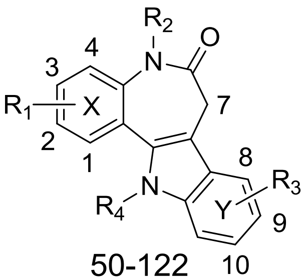
Group	Data set	Chemotype	Methods	QSAR results
Kunick et al.[64]	training set (52) test set (21)	Paullones	CoMSIA	best CoMSIA training set R^2 = 0.871, LOO-CV R^2 = 0.554
Lescot et al.[180]	74	3-anilino-4-arylmaleimides	CoMFA	best CoMFA training set R^2 = 0.891, LOO-CV R^2 = 0.805
Zeng et al.[77]	training set (30) test set (5)	aloisines	CoMFA CoMSIA	best CoMFA training set R^2 = 0.917, LOO-CV R^2 = 0.584, best CoMSIA training set R^2 = 0.938, LOO-CV R^2 = 0.673
Katritzky et al.[74]	training set (187) validation set (90)	3-anilino-4-arylmaleimides, pyrozolopyridazines, pyrazolopyridines, pyrazolopyrimidines	ANN	training set R^2 = 0.782, validation set R^2 = 0.679
Zhang et al.[78]	training set (34) test set (8)	indirubins	CoMFA CoMSIA	receptor-based CoMSIA training set R^2 = 0.908, LOO-CV R^2 = 0.766
Dessalew et al.[181]	training set (49) test set (12)	pyrazolopyrimidines	CoMFA CoMSIA	best CoMFA training set R^2 = 0.98, LOO-CV R^2 = 0.53, test set R^2 = 0.47 best CoMSIA training set R^2 = 0.92, LOO-CV R^2 = 0.48 test set R^2 = 0.48
Lather et al.[182]	training set (36) test set (8)	indirubins	multiple linear correlation for 2D QSAR, atom-based PHASE 3D QSAR	2D QSAR training set R^2 = 0.93, test set R^2 = 0.60 3D QSAR training set R^2 = 0.97,

				test set R^2 = 0.91
Taha et al.[80]	training set (123) test set (29)	3-anilino-4-arylmaleimides, pyrozolopyridazines, pyrazolopyridines	GFA-MLR-QSAR	training set R^2 = 0.663, LOO-CV R^2 = 0.592, test set R^2 = 0.695
Prasanna et al.[183]	training set (56) test set (18)	3-anilino-4-arylmaleimides	CoMFA CoMSIA	best CoMFA training set R^2 = 0.942, LOO-CV R^2 = 0.844, test set R^2 = 0.779 best CoMSIA training set R^2 = 0.932, LOO-CV R^2 = 0.833 test set R^2 = 0.803
Goodarzi et al.[81]	training set (123) test set (29)	3-anilino-4-arylmaleimides, pyrozolopyridazines, pyrazolopyridines	linear/nonlinear regression methods	best model obtained using SVM combining with fuzzy rough set ACO as variable selection method training set R^2 = 0.960, test set R^2 = 0.927
Fang et al.[83]	training set (30) test set (8)	Benzofuran-3-yl-(indol-3-yl)maleimides	CoMFA CoMSIA	best CoMFA training set R^2 = 0.984, LOO-CV R^2 = 0.602, test set R^2 = 0.905 best CoMSIA training set R^2 = 0.983, LOO-CV R^2 = 0.665 test set R^2 = 0.761

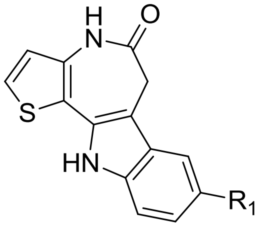
Appendix: B. Structures and bioactivities of the compounds in Group I.



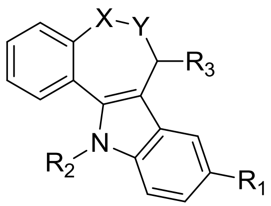
1-49
175-187[b]



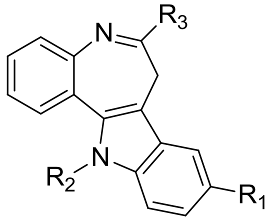
50-122
188-203[b]



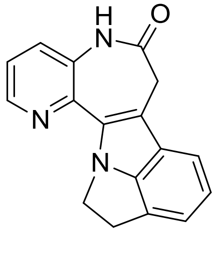
123-125
204-205[b]



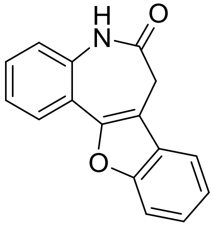
126-129
206[b]



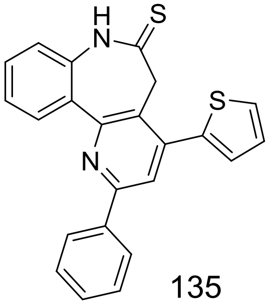
130-132



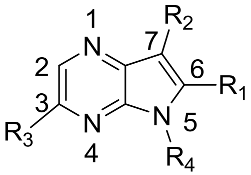
133



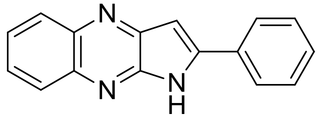
134



135



136-173
207-214[b]



174

No.	R1[a]	R2[a]	R3[a]	R4[a]	X[a]	Y	pIC50				
							Expt.	Models			
								I	II	III	IV
1	6-Br	H	H	–	$NO(CH_2)_2$Piperazinyl	–	8.481	8.402	8.258	8.119	8.102
2	5,6-diCl	H	H	–	NOH	–	8.398	8.318	8.174	8.082	8.040
3	5,6-diCl	H	H	–	NOAc	–	8.398	8.318	8.174	8.149	8.155
4	6-Br	H	H	–	NOH	–	8.301	8.221	8.077	7.881	7.766
5	6-Br	H	H	–	$NO(CH_2)_2$Piperazinyl-N-EtOH	–	8.301	8.330	8.077	8.119	8.088
6	5-Me-6-Br	H	H	–	NOH	–	8.222	8.302	7.998	7.931	7.962
7	5-NO_2-6-Br	H	H	–	NOAc	–	8.222	8.302	7.998	7.989	7.997
8	5-Me-6-Br	H	H	–	NOAc	–	8.155	8.075	7.931	8.086	7.980
9	5-NO_2-6-Br	H	H	–	NOH	–	8.155	8.075	7.931	7.963	7.871
10	6-Br	H	H	–	$NO(CH_2)_2$Piperazinyl-	–	8.155	8.075	7.931	8.090	8.076

					N-Me						
11	5-I	H	H	–	NOH	–	8.046	7.966	7.822	7.852	7.632
12	6-Br	H	H	–	NOAc	–	8.000	7.920	7.776	7.925	7.860
13	6-I	H	H	–	NOAc	–	7.886	7.927	7.662	7.816	7.710
14	6-Br	H	H	–	NO(CH_2) $_2$Piperazinyl-N-EtOEtOH	–	7.854	7.774	7.630	7.847	7.716
15	6-Cl	H	H	–	NOAc	–	7.770	7.747	7.546	7.781	7.743
16	H	H	H	–	NOH	–	7.658	7.578	7.434	7.325	7.266
17	5-Me-6-Br	H	H	–	O	–	7.602	7.682	7.378	7.192	7.256
18	6-Br	H	H	–	$NO(CH_2)_2N(CH_2)_4$	–	7.585	7.665	7.361	7.538	7.542
19	5,6-diCl	H	H	–	O	–	7.523	7.443	7.299	7.294	7.233
20	6-Br	H	H	–	$NO(CH_2)_2NMe_2$	–	7.481	7.408	7.258	7.491	7.490
21	6-Br	H	H	–	$NOCH_2CH(OH)CH_2OH$	–	7.469	7.389	7.244	7.461	7.472
22	6-Br	H	H	–	$NO(CH_2)_2NEt_2$	–	7.456	7.536	7.232	7.491	7.418
23	5-NO_2	H	H	–	O	–	7.377	7.297	7.153	7.234	7.242
24	6-Br	H	H	–	O	–	7.347	7.266	7.123	7.116	7.031
25	6-I	H	H	–	O	–	7.260	7.339	7.036	7.109	7.061
26	5-Br	H	H	–	O	–	7.260	7.180	7.036	7.129	7.094
27	6-CH=CH_2	H	H	–	NOH	–	7.222	7.302	6.998	7.299	7.156
28	6-Br	H	H	–	NO(CH_2) $_2$Morpholinyl	–	7.222	7.302	6.998	7.498	7.447
29	5-Me	H	H	–	O	–	7.208	7.128	6.984	6.982	6.891
30	6-CH=CH_2	H	H	–	NOAc	–	7.187	7.268	6.963	7.277	7.269
31	6-Br	H	H	–	$NO(CH_2)_2$NMe-2,3-diOH-*n*Pr	–	7.174	7.094	6.949	7.441	7.408
32	5-I	H	H	–	O	–	7.167	7.247	6.943	7.142	7.019
33	5-F	H	H	–	O	–	7.108	7.028	6.884	7.030	6.937
34	5-NO_2-6-Br	H	H	–	O	–	7.000	7.080	6.776	7.132	7.136
35	6-Br	6’-Br	H	–	NOH	–	6.921	7.001	6.697	7.082	7.079
36	6-Cl	H	H	–	O	–	6.854	6.934	6.630	6.896	6.757
37	6-Br	H	H	–	$NO(CH_2)_2Br$	–	6.854	6.934	6.630	7.013	7.069
38	H	H	H	–	$NOCH_3$	–	6.824	6.904	6.600	6.599	6.710
39	H	H	H	–	NOAc	–	6.699	6.779	6.475	6.986	6.950

40	H	6’-Br	H	–	NOH	–	6.469	6.549	6.444	7.009	6.872
41	H	5’-Br	H	–	O	–	6.456	6.376	6.232	6.251	6.384
42	H	H	H	–	O	–	6.000	6.080	6.198	6.049	6.101
43	6-Br	6’-Br	H	–	O	–	5.347	5.427	5.571	5.742	5.815
44	H	6’-Br	H	–	O	–	4.658	4.738	4.882	5.627	5.512
45	6-Br	H	Me	–	O	–	4.000	4.080	4.224	4.411	4.448
46	6-Br	H	Me	–	NOH	–	4.000	4.080	4.224	5.057	5.103
47	H	H	Me	–	O	–	4.000	3.920	4.224	4.083	4.223
48	H	H	Me	–	NOH	–	4.000	4.080	4.224	4.591	4.728
49	4-Cl	H	H	–	O	–	4.000	4.080	4.224	4.824	4.887
50	H	H	9-NO_2	H	–	–	8.398	8.318	8.174	7.831	7.720
51	H	H	9-CN	H	1-N	–	8.097	8.017	7.873	7.377	7.473
52	H	H	9-CF_3	H	1-N	–	8.097	8.017	7.873	7.289	7.220
53	H	H	9-CN	H	–	–	8.000	7.920	7.776	7.391	7.520
54	2,3-diOMe	H	9-NO_2	H	–	–	7.886	7.807	7.662	7.569	7.606
55	H	H	9-Cl	H	2-N	–	7.745	7.664	7.521	7.180	7.275
56	2,3-diOMe	H	9-CN	H	–	–	7.745	7.665	7.521	7.382	7.444
57	3-OH	H	9-Br	H	–	–	7.745	7.665	7.521	7.039	7.214
58	H	H	9-Br	H	–	–	7.638	7.652	7.414	7.272	7.362
59	H	H	9-Cl	H	–	–	7.620	7.540	7.395	7.304	7.235
60	H	H	9-I	H	1-N	–	7.602	7.522	7.378	7.148	7.188
61	2-CN	H	9-CN	H	–	–	7.553	7.633	7.329	7.283	7.219
62	H	H	9-CF_3	H	–	–	7.523	7.443	7.299	7.195	7.133
63	3-$O(CH_2)_4NH_2$	H	9-Br	H	–	–	7.523	7.443	7.299	7.162	7.200
64	2-CH_2CH_2CN	H	9-CF_3	H	–	–	7.481	7.401	7.257	7.147	7.100
65	H	H	9-F	H	2-N	–	7.292	7.373	7.068	7.023	6.953
66	H	H	9-Br	H	2-N	–	7.284	7.364	7.060	7.121	7.089
67	H	H	9-Cl	H	1-N		7.201	7.280	6.977	6.760	6.805
68	3-OMe	H	9-NO_2	H	–	–	7.155	7.075	6.931	7.253	7.251
69	H	H	9-Br	CH_2CO_2-Me	–	–	7.125	7.045	6.901	6.665	6.628
70	2,3-diOMe	H	9-CF_3	H	–	–	7.125	7.045	6.901	6.857	6.913
71	H	H	9-F	H	–	–	7.097	7.017	6.873	6.988	6.866

72	2-CH_2COMe	H	9-CF_3	H	–	–	7.046	6.966	6.822	6.804	6.786
73	2,3-diOMe	H	9-Br	H	–	–	7.000	6.920	6.776	6.712	6.714
74	H	H	9-Me	H	1-N	–	6.886	6.966	6.662	6.592	6.576
75	H	H	9-Me	H	–	–	6.886	6.806	6.662	6.596	6.567
76	H	H	10-Br	H	–	–	6.854	6.774	6.630	6.880	6.736
77	2-Br	H	H	H	–	–	6.699	6.779	6.475	6.743	6.552
78	2-Br	H	9-NO_2	H	–	–	6.699	6.779	6.475	6.851	6.708
79	3-OMe	H	9-CF_3	H	–	–	6.620	6.700	6.396	6.700	6.696
80	2-I	H	H	H	–	–	6.602	6.522	6.378	6.535	6.470
81	H	H	9-Br	$(CH_2)_2OH$	–	–	6.523	6.443	6.299	6.275	6.378
82	H	H	9-OMe	H	2-N	–	6.409	6.329	6.186	6.347	6.302
83	H	H	H	Me	2-N	–	6.398	6.317	6.194	6.129	6.205
84	H	H	9-Br	Me	–	–	6.398	6.478	6.251	6.594	6.550
85	2-CH=CHCN	H	9-CF_3	H	–	–	6.398	6.393	6.193	6.248	6.339
86	H	H	H	Bn	2-N	–	6.387	6.307	6.193	5.980	5.972
87	H	H	11-Me	H			6.301	6.221	6.257	6.203	6.225
88	H	H	H	H	–	–	6.208	6.288	6.154	6.059	6.030
89	H	H	8,10-diCl	H	1-N	–	6.097	6.177	6.180	6.007	5.979
90	2-Br	H	9-Br	H	–	–	6.097	6.017	6.193	6.176	6.063
91	H	H	11-Br	H	–	–	6.046	6.126	6.225	6.210	6.185
92	2,3-diOMe	H	H	H	–	–	6.046	5.966	6.192	6.134	6.105
93	2-CH=$CHCO_2$-Me	H	9-CF_3	H	–	–	6.046	6.126	6.190	6.202	6.197
94	H	H	H	Me	1-N	–	6.000	5.920	6.132	5.873	5.915
95	2-CN	H	H	H	–	–	5.886	5.966	6.110	6.401	6.247
96	H	H	9-OH	H	1-N		5.854	5.849	6.078	5.987	5.988
97	2-CH=CHCOMe	H	9-CF_3	H	–	–	5.854	5.774	6.078	6.116	6.106
98	H	H	9,11-diF	H	1-N	–	5.745	5.665	5.969	6.020	5.960
99	2-Br	H	9-CF_3	H	–	–	5.699	5.779	5.923	5.928	5.933
100	2-CH=CH-1-	H	9-CF_3	H	–	–	5.699	5.779	5.923	5.902	5.962

	OH-*cy*He										
101	H	Me	9-Br	H	–	–	5.678	5.598	5.902	5.629	5.655
102	H	H	9-OMe	H	–	–	5.658	5.738	5.882	5.709	5.732
103	2-I	H	9-CF_3	H	–	–	5.658	5.578	5.881	5.775	5.927
104	H	H	9-Br	CO_2*t*Bu	–	–	5.638	5.718	5.862	5.654	5.616
105	2-Br	H	9-CN	H	1-N	–	5.420	5.500	5.644	6.226	6.058
106	H	H	9-Br	Allyl	–	–	5.398	5.478	5.622	5.536	5.613
107	2-I	H	9-Br	H	–	–	5.377	5.457	5.601	5.777	5.759
108	4-OH	H	9-Br	H	–	–	5.367	5.447	5.591	5.535	5.744
109	H	H	8,10-diCl	H	–	–	5.301	5.381	5.525	5.601	5.620
110	H	H	9-CF_3	H	4-N		5.276	5.356	5.500	5.976	5.930
111	H	H	H	H	–	11-N	5.260	5.179	5.484	5.322	5.502
112	H	H	H	Ph	1-N		5.000	5.080	5.224	5.248	5.299
113	H	Bn	9-Br	H	–	–	5.000	4.976	5.224	4.728	4.845
114	H	H	9-NH_2	H	–	–	4.921	5.001	5.145	5.591	5.494
115	H	H	H	Et	1-N	–	4.886	4.966	5.110	5.328	5.321
116	4-OMe	H	9-Br	H	–	–	4.796	4.876	5.019	5.160	5.341
117	H	Et	9-Br	H	–	–	4.620	4.700	4.844	5.058	4.954
118	H	H	9-NHAc	H	–	–	4.387	4.467	4.611	5.321	5.309
119	H	H	9-CO_2H	H	1-N	–	4.000	4.080	4.224	5.181	5.277
120	4-OMe	H	H	H	–	–	3.854	3.934	4.078	4.672	4.695
121	H	H	9-Cl	H	4-N	–	3.301	3.381	3.525	4.834	4.947
122	H	CO_2*t*Bu	9-Br	CO_2*t*Bu	–	–	3.194	3.274	3.418	3.530	3.681
123	CN	–	–	–	–	–	6.481	6.401	6.258	6.464	6.535
124	Cl	–	–	–	–	–	6.398	6.318	6.192	6.157	6.208
125	Me	–	–	–	–	–	5.886	5.966	6.110	6.014	6.032
126	NO_2	H	H	–	NH	C=S	6.222	6.202	6.212	6.141	6.173
127	Br	H	CO_2*t*B	–	N CO_2*t*Bu	C=	3.886	3.966	4.110	4.041	4.103

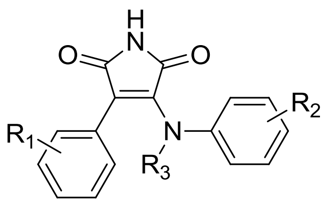
			u			O					
128	Br	H	H	–	CH_2	CH 2	3.745	3.825	3.969	4.300	4.424
129	Br	CO_2*t*Bu	CO_2*t*B u	–	N CO_2*t*Bu	C= O	3.301	3.381	3.525	3.541	3.660
130	Br	H	NHOH	–	–	–	6.125	6.045	6.195	5.973	6.067
131	Br	H	SMe	–	–	–	5.921	5.841	6.145	5.672	5.651
132	NO_2	Me	SMe	–	–	–	3.000	3.080	3.224	4.025	3.988
133	–	–	–	–	–	–	6.387	6.307	6.191	6.072	5.998
134	–	–	–	–	–	–	5.000	5.080	5.224	5.056	5.177
135	–	–	–	–	–	–	3.456	3.536	3.680	3.846	3.857
136	4-OMe-Ph	$(CH_2)_2Me$	H	H	–	–	6.398	6.478	6.190	6.214	6.241
137	4-OMe-Ph	Me	H	H	–	–	6.337	6.257	6.193	6.135	6.139
138	4-OMe-Ph	$CHMe_2$	H	H	–	–	6.301	6.274	6.193	6.124	6.139
139	4-OSO_2NMe_2-Ph	Me	H	H	–	–	6.301	6.221	6.193	5.768	5.894
140	4-OH-Ph	Me	H	H	–	–	6.284	6.364	6.192	6.140	6.116
141	4-OH-Ph	$(CH_2)_3Me$	H	H	–	–	6.187	6.107	6.191	6.024	6.054
142	4-Cl-Ph	$CHMe_2$	H	H	–	–	6.125	6.045	6.193	5.850	5.849
143	3-Thienyl	H	H	H	–	–	6.097	6.017	6.193	5.881	5.761
144	4-OMe-Ph	$(CH_2)_3Me$	H	H	–	–	6.036	5.956	6.193	5.984	5.940
145	Ph	Bn	H	H	–	–	6.000	5.920	6.193	5.623	5.582
146	$C(CH_2)_2$-4-Cl-Ph	H	H	H	–	–	6.000	5.920	6.152	5.703	5.737
147	4-OMe-Ph	CH_2*cy*Pr	H	H	–	–	5.959	6.038	6.183	5.841	5.860
148	4-OH-Ph	H	H	H	–	–	5.921	6.001	6.145	5.860	5.838
149	2-Thienyl	H	H	H	–	–	5.921	5.841	6.144	5.734	5.666
150	4-Cl-Ph	Me	H	H	–	–	5.770	5.689	5.994	5.632	5.568
151	4-F-Ph	H	H	H	–	–	5.721	5.641	5.945	5.570	5.601
152	3,4-diOMe-Ph	Me	H	H	–	–	5.699	5.619	5.923	5.619	5.610
153	Ph	H	H	H	–	–	5.638	5.558	5.862	5.380	5.393
154	4-OMe-Ph	$(CH_2)_3Cl$	H	H	–	–	5.602	5.682	5.826	5.792	5.745
155	4-OH-Ph	CH_2*cy*Pr	H	H	–	–	5.523	5.603	5.747	5.738	5.703

156	3-OMe-Ph	H	H	H	–	–	5.495	5.415	5.719	5.391	5.372
157	2-OMe-Ph	H	H	H	–	–	5.481	5.437	5.705	5.605	5.544
158	4-CN-Ph	H	H	H	–	–	5.319	5.399	5.543	5.400	5.334
159	4-Cl-Ph	$(CH_2)_3Me$	H	H	–	–	5.229	5.309	5.453	5.425	5.321
160	4-Br-Ph	H	H	H	–	–	5.222	5.302	5.446	5.167	5.247
161	4-OMe-Ph	CH_2*cy*He	H	H	–	–	5.167	5.087	5.392	5.236	5.185
162	4-CF_3-Ph	H	H	H	–	–	5.143	5.223	5.367	5.206	5.247
163	4-Cl-Ph	CH_2*cy*He	H	H	–	–	5.097	5.093	5.321	5.148	5.120
164	4-OMe-Ph	$(CH_2)_6Me$	H	H	–	–	5.000	5.080	5.224	4.858	4.908
165	4-NMe_2-Ph	H	H	H	–	–	4.921	5.001	5.145	5.062	5.084
166	2-Furanyl	H	H	H	–	–	4.824	4.904	5.048	5.063	5.065
167	2-Pyridyl	H	H	H	–	–	4.824	4.904	5.048	4.980	4.990
168	$C(CH_2)_2$-4-Cl-Ph	Me	H	H	–	–	4.745	4.825	4.969	5.204	5.122
169	3,5-diOMe-Ph	H	H	H	–	–	4.222	4.302	4.446	4.471	4.521
170	3,4,5-triOMe-Ph	H	H	H	–	–	4.071	4.150	4.295	4.217	4.352
171	3,5-diCl-Ph	H	H	H	–	–	4.000	4.080	4.224	4.588	4.511
172	Ph	H	Me	H	–	–	4.000	4.080	4.224	4.373	4.385
173	Ph	H	H	Me	–	–	4.000	4.080	4.224	4.586	4.488
174	–	–	–	–	–	–	4.000	3.920	4.224	4.274	4.243
175[b]	6-I	H	H	–	NOH	–	8.000	8.052	6.194	7.814	7.476
176[b]	6-Br	H	H	–	$NO(CH_2)_2$Piperazinyl-N-EtOMe	–	7.959	7.778	6.230	7.891	7.837
177[b]	6-Cl	H	H	–	NOH	–	7.699	8.200	6.204	7.589	7.523
178[b]	6-Br	H	H	–	$NOCH_3$	–	7.523	7.526	6.148	6.962	7.282
179[b]	6-Br	H	H	–	$NO(CH_2)_2OH$	–	7.523	7.320	6.228	7.302	7.490
180[b]	6-Br	H	H	–	$NOCONEt_2$	–	7.523	6.866	6.193	7.613	7.629
181[b]	6-Br	H	H	–	$NO(CH_2)_2N(EtOH)_2$	–	7.398	6.631	6.197	7.592	7.431
182[b]	5-Cl	H	H	–	O	–	7.301	6.740	6.291	6.937	6.855
183[b]	6-F	H	H	–	NOAc	–	7.046	7.150	6.199	7.826	7.417
184[b]	6-F	H	H	–	NOH	–	6.886	8.092	6.200	7.793	7.35
185[b]	6-$CH=CH_2$	H	H	–	O	–	6.620	6.308	6.198	6.703	6.641

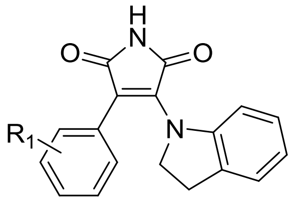
186[b]	5-Br	5’-Br	H	–	O	–	6.602	6.114	6.193	6.108	6.367
187[b]	6-F	H	H	–	O	–	6.187	6.851	6.447	6.961	6.729
188[b]	H	H	9-CF_3	H	2-N	–	7.886	7.016	6.086	6.843	6.669
189[b]	H	H	9-CN	H	2-N	–	7.678	8.602	6.348	6.976	7.094
190[b]	2-OMe	H	9-NO_2	H	–	–	7.658	7.051	6.223	6.980	7.108
191[b]	H	H	9-F	H	1-N	–	7.097	6.582	6.236	6.687	6.204
192[b]	H	H	9-CN-11-I	H	1-N	–	6.921	7.703	6.193	6.546	6.599
193[b]	2,3-diOH	H	9-Br	H	–	–	6.921	6.501	6.314	6.664	6.719
194[b]	3-OMe	H	9-CN	H	–	–	6.886	6.912	6.202	7.078	7.065
195[b]	H	H	11-Cl	H	–	–	6.699	6.512	6.343	6.395	6.195
196[b]	2-C≡CH_2OH	H	9-CF_3	H	–	–	6.699	6.474	6.194	6.474	6.488
197[b]	H	CH_2CO_2Me	9-Br	H	–	–	6.301	4.542	6.193	4.989	5.083
198[b]	H	H	11-Et	H	–	–	6.155	6.423	6.197	6.110	5.972
199[b]	H	H	H	H	1-N	–	6.097	6.874	6.193	6.117	6.029
200[b]	H	H	9-OMe	H	1-N	–	6.097	5.204	6.176	5.812	5.853
201[b]	H	H	H	Et	2-N	–	6.097	5.661	6.157	5.918	5.926
202[b]	H	H	9-Br	Et	–	–	5.824	5.980	6.152	5.705	5.759
203[b]	H	H	H	Bn	1-N	–	5.678	5.501	6.190	5.574	5.527
204[b]	NO_2	–	–	–	–	–	7.456	7.063	6.193	6.769	6.557
205[b]	Br	–	–	–	–	–	6.921	6.671	6.193	6.192	6.369
206[b]	Br	H	H	–	NH	C=S	5.699	5.773	6.195	5.796	5.845
207[b]	4-OMe-Ph	$CH_2CH=CH_2$	H	H	–	–	6.222	6.352	6.192	5.980	6.013
208[b]	4-OMe-Ph	H	H	H	–	–	5.959	5.304	6.191	5.674	5.589
209[b]	4-OH-Ph	$(CH_2)_2Me$	H	H	–	–	5.745	5.994	6.191	6.033	6.027
210[b]	4-Me-Ph	H	H	H	–	–	5.585	5.055	6.045	5.290	5.151
211[b]	2-OH-Ph	H	H	H	–	–	5.187	5.299	6.193	5.761	5.693
212[b]	4-Cl-Ph	Bn	H	H	–	–	5.167	5.524	6.193	5.273	5.292
213[b]	4-(2-1,3-dioxolano)-Ph	H	H	H	–	–	4.699	5.030	6.193	4.832	5.074
214[b]	1-Naphthyl	H	H	H	–	–	4.569	6.308	6.193	5.398	5.330

[a] Ac = Acetyl, Br = Bromo, Bu = Butyl, Bn = Benzyl, Cl = Chloro, Et = Ethyl, Me = Methyl, Ph = Phenyl, Pr = Propyl, *n* = *normal*, *cy* = *cyclo*, *t* = *tert*; [b] Test set.

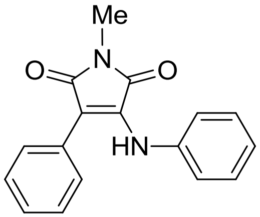
Appendix: C. Structures and bioactivities of the compounds in Group II.



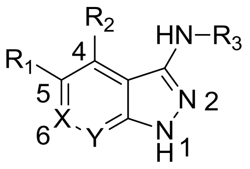
1-54
129-141^b



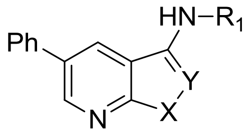
55-58
142-144^b



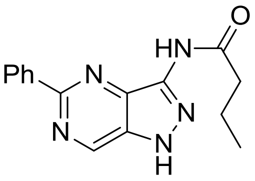
59



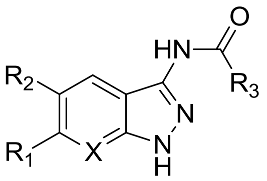
60-88
145-150^b



89-92



93



94-128
151-157^b

No.	R$_1$^{a}	R$_2$^{a}	R$_3$^{a}	X^a	Y	pIC$_{50}$				
						Expt.	Model			
							I	II	III	IV
1	3-NO$_2$	3,5,-di-Cl-4-OH	H	–	–	7.699	7.597	7.575	7.414	7.456
2	2-NO$_2$	4-Cl-3-CO$_2$H	H	–	–	7.553	7.451	7.429	7.286	7.445
3	3-Cl	3,5,-di-Cl-4-OH	H	–	–	7.237	7.135	7.113	7.140	7.149
4	3-NO$_2$	3-Cl-4-OH	H	–	–	7.229	7.170	7.105	7.197	7.202
5	3-NO$_2$	3-Cl	H	–	–	7.155	7.053	7.031	6.918	6.984
6	4-NO$_2$	3,5,-di-Cl-4-OH	H	–	–	7.149	7.251	7.173	7.183	7.173
7	2-Cl	4-Cl-3-CO$_2$H	H	–	–	7.131	7.031	7.007	7.001	7.078
8	3-Cl	4-Cl-3-CO$_2$H	H	–	–	7.119	7.205	6.995	7.071	7.104
9	3-NO$_2$	3-CO$_2$H	H	–	–	7.102	7.084	6.979	6.938	7.026
10	2-OMe	3,5,-di-Cl-4-OH	H	–	–	7.086	6.985	6.962	7.039	7.075
11	3-OMe	4-Cl-3-CO$_2$H	H	–	–	7.071	7.043	6.946	7.081	7.102
12	4-Cl	3,5,-di-Cl-4-OH	H	–	–	7.041	7.143	7.028	7.014	7.043

13	2-Cl	3,5,-di-Cl-4-OH	H	–	–	7.032	7.134	6.952	6.923	7.111
14	2-NO_2	3-Cl	H	–	–	6.983	7.085	6.885	6.878	6.916
15	2-NO_2	3-Cl-4-OH	H	–	–	6.983	7.085	7.027	7.002	7.031
16	2-OMe	4-SMe	H	–	–	6.959	7.060	6.835	6.769	6.814
17	2-OMe	3-Cl	H	–	–	6.943	7.045	6.819	6.777	6.832
18	3-NO_2	4-OH	H	–	–	6.910	6.808	6.849	6.820	6.795
19	3-Cl	3-CO_2H	H	–	–	6.873	6.900	6.760	6.860	6.843
20	2-Cl	3-CO_2H	H	–	–	6.866	6.895	6.810	6.818	6.856
21	3-OMe	3,5,-di-Cl-4-OH	H	–	–	6.848	6.950	6.972	6.981	6.984
22	H	4-Cl-3-CO_2H	H	–	–	6.845	6.947	6.823	6.847	6.816
23	H	3,5,-di-Cl-4-OH	H	–	–	6.827	6.752	6.842	6.845	6.831
24	2-Cl	3-Cl-4-OH	H	–	–	6.818	6.811	6.847	6.794	6.846
25	3-NO_2	4-SMe	H	–	–	6.818	6.716	6.751	6.768	6.772
26	4-OMe	3-Cl	H	–	–	6.807	6.705	6.683	6.713	6.679
27	2-Cl	4-SMe	H	–	–	6.793	6.752	6.674	6.758	6.711
28	4-Cl	3-Cl-4-OH	H	–	–	6.762	6.743	6.757	6.798	6.766
29	4-Cl	3-CO_2H	H	–	–	6.730	6.833	6.798	6.745	6.751
30	2-Cl	3-Cl	H	–	–	6.710	6.644	6.633	6.629	6.643
31	3-OMe	3-CO_2H	H	–	–	6.710	6.639	6.748	6.705	6.724
32	4-OMe	3-CO_2H	H	–	–	6.670	6.772	6.794	6.672	6.689
33	2-Cl	H	H	–	–	6.666	6.563	6.561	6.513	6.580
34	2-OMe	H	H	–	–	6.666	6.768	6.542	6.532	6.522
35	3-NO_2	3-OH	H	–	–	6.627	6.568	6.738	6.614	6.656
36	4-OMe	4-SMe	H	–	–	6.614	6.716	6.738	6.668	6.598
37	2-NO_2	3-OH	H	–	–	6.600	6.679	6.724	6.654	6.672
38	3-OMe	3-Cl	H	–	–	6.590	6.692	6.532	6.658	6.574
39	H	3-CO_2H	H	–	–	6.536	6.522	6.633	6.530	6.542
40	4-Cl	4-OH	H	–	–	6.499	6.464	6.500	6.444	6.471
41	2-Cl	3-OH	H	–	–	6.427	6.325	6.551	6.407	6.465
42	4-OMe	H	H	–	–	6.409	6.511	6.424	6.338	6.375
43	4-NO_2	4-SMe	H	–	–	6.407	6.508	6.530	6.546	6.561
44	4-Cl	3-Cl	H	–	–	6.350	6.452	6.474	6.372	6.407
45	3-OMe	3-OH	H	–	–	6.326	6.224	6.432	6.386	6.371

46	4-OMe	3-OH	H	–	–	6.318	6.420	6.442	6.370	6.392
47	4-Cl	H	H	–	–	6.289	6.391	6.413	6.284	6.285
48	H	H	H	–	–	6.277	6.358	6.400	6.225	6.273
49	4-Cl	4-SMe	H	–	–	6.277	6.360	6.400	6.389	6.370
50	3-Cl	4-SMe	H	–	–	6.274	6.376	6.398	6.393	6.412
51	H	3-OH	H	–	–	6.152	6.255	6.277	6.249	6.202
52	3-NO$_2$	H	Me	–	–	5.854	5.956	5.978	6.098	6.143
53	3-Cl	3-OH	H	–	–	5.830	5.933	5.954	6.165	6.084
54	3-OMe	H	Me	–	–	5.347	5.449	5.471	5.681	5.774
55	2-NO$_2$	–	–	–	–	6.883	6.781	6.827	6.712	6.815
56	2-OMe	–	–	–	–	6.728	6.626	6.722	6.458	6.581
57	2-Cl	–	–	–	–	6.472	6.371	6.570	6.488	6.552
58	4-Cl	–	–	–	–	5.850	5.952	5.974	6.110	6.120
59	–	–	–	–	–	5.301	5.403	5.425	5.703	5.727
60	Ph	H	CO*n*Pr	N	N	8.398	8.500	8.274	8.039	7.963
61	Ph	H	CO*n*Bu	N	N	8.398	8.296	8.274	8.255	8.191
62	Ph	H	CO(CH$_2$)$_3$Morpholinyl	N	N	8.301	8.199	8.177	7.839	7.932
63	2,3-diF-Ph	H	CO(CH$_2$)$_3$NMe$_2$	N	N	8.301	8.199	8.177	8.197	8.104
64	Ph	H	CO*cy*Pent	CH	N	8.301	8.199	8.177	8.096	8.108
65	Ph	H	CO*n*Pr	N	CH	8.155	8.052	8.031	7.912	7.863
66	Ph	H	CO(CH$_2$)$_3$Piperazinyl-N-Et	N	N	8.155	8.053	8.031	8.101	8.055
67	2,3-diF-Ph	H	CO*n*Pr	CH	N	8.155	8.053	8.031	7.916	7.919
68	Ph	H	CO(CH$_2$)$_3$Pyrrolidine	N	N	7.959	8.061	7.835	8.049	7.966
69	3-Pyridyl	H	CO*n*Pr	CH	N	7.959	7.856	7.835	7.469	7.488
70	Ph	H	CO*i*Pr	CH	N	7.721	7.794	7.597	7.392	7.391
71	3-F-Ph	H	CO*n*Pr	CH	N	7.699	7.801	7.601	7.498	7.526
72	Ph	H	CO(CH$_2$)$_3$NMe$_2$	N	N	7.658	7.760	7.534	7.863	7.789
73	2-Cl-Ph	H	CO*n*Pr	CH	N	7.569	7.540	7.445	7.511	7.458
74	Ph	H	COEt	CH	N	7.367	7.264	7.243	7.134	7.004
75	Ph	H	CO*n*Pr	CH	N	7.252	7.300	7.376	7.291	7.257
76	Ph	H	CO*n*Pr	CH	CH	7.004	7.106	7.129	6.940	6.936
77	2-Naphthyl	H	CO*n*Pr	CH	N	6.772	6.874	6.896	6.961	6.958
78	Ph	Ph	H	N	N	6.602	6.704	6.726	6.536	6.470

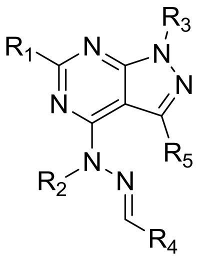
79	B(OCMe$_2$CMe$_2$O)	H	CO*n*Pr	CH	N	6.449	6.522	6.573	6.887	6.901
80	Ph	H	H	CH	N	6.367	6.264	6.243	6.154	6.146
81	4-Pyridyl	H	CO*n*Pr	CH	N	6.354	6.456	6.478	6.638	6.630
82	Ph	H	H	N	N	6.276	6.318	6.390	6.160	6.171
83	Ph	Ph	CO*n*Pr	N	N	6.161	6.262	6.284	6.458	6.483
84	4-Ph-Ph	H	CO*n*Pr	CH	N	6.070	6.088	6.194	6.606	6.614
85	Ph	H	H	N	CH	5.900	5.923	6.024	6.030	5.964
86	H	H	CO*n*Pr	CH	N	5.630	5.704	5.754	6.006	5.952
87	Ph	H	CONHEt	CH	N	5.551	5.653	5.675	6.193	6.171
88	Ph	H	SO$_2$Me	CH	N	5.447	5.549	5.571	5.862	5.923
89	COMe	–	–	NH	N	6.536	6.638	6.660	6.410	6.421
90	H	–	–	NMe	N	5.301	5.403	5.425	5.623	5.520
91	H	–	–	O	N	5.301	5.403	5.425	5.702	5.668
92	H	–	–	NH	CH	5.301	5.403	5.426	5.551	5.511
93	–	–	–	–	–	5.569	5.671	5.693	6.174	6.215
94	4-HO-Ph	Br	CO*cy*Pr	N	–	9.097	8.995	8.973	8.402	8.363
95	4-HO-Ph	Cl	CO*cy*Pr	N	–	9.000	8.898	8.876	8.499	8.573
96	4-HO-Ph	Br	CO-4-Piperidine-N-Me	N	–	9.000	8.898	8.876	8.555	8.649
97	4-HO-Ph	Br	CO(CH$_2$)$_3$Piperazinyl-N-Et	N	–	8.398	8.473	8.274	8.270	8.265
98	3-Br-4-HO-Ph	H	CO*cy*Pr	N	–	8.301	8.199	8.177	8.194	8.168
99	4-HO-Ph	Me	CO*cy*Pr	N	–	8.222	8.324	8.098	7.862	7.817
100	2-Thienyl	Br	CO*cy*Pent	N	–	8.155	8.159	8.031	7.912	7.876
101	2-Furanyl	Br	CO*cy*Pr	N	–	8.155	8.053	8.031	7.858	7.821
102	4-HO-Ph	H	CO*cy*Pr	N	–	8.097	8.121	7.973	7.903	7.901
103	3,4-diHO-Ph	H	CO*cy*Pr	N	–	8.097	8.199	7.973	8.032	8.045
104	3-HO-Ph	H	CO*cy*Pr	N	–	7.921	8.023	7.797	7.826	7.825
105	2-Furanyl	Br	CO-(±)-3-Pyrrolidine-N-Bn	N	–	7.854	7.860	7.730	7.884	7.837
106	4-HO-Ph	H	CO*cy*Pr	CH	–	7.824	7.722	7.700	7.149	7.150
107	2-Thienyl	Br	CO-CH$_2$-4-Piperidine-N-Et	N	–	7.745	7.643	7.621	7.745	7.684
108	4-HO-Ph	Ph	CO*cy*Pr	N	–	7.620	7.722	7.496	7.527	7.528
109	2-Thienyl	Br	CO*cy*Pr	N	–	7.409	7.402	7.285	7.410	7.378
110	5-Indolyl	H	CO*cy*Pr	CH	–	7.377	7.275	7.252	7.183	7.217
111	2-Furanyl	H	CO*cy*Pr	CH	–	7.301	7.199	7.177	7.137	7.059

112	Ph	Br	CO*cy*Pr	N	–	7.125	7.023	7.249	7.122	7.146
113	Ph	CN	CO*cy*Pr	N	–	7.060	6.959	6.990	7.122	7.091
114	2-Thiazolyl	Br	CO*cy*Pr	N	–	7.004	7.107	7.080	7.269	7.211
115	3-MeO-Ph	H	CO*cy*Pr	N	–	6.903	6.454	6.779	6.695	6.794
116	2-Thienyl	H	CO*cy*Pr	CH	–	6.668	6.753	6.792	6.888	6.750
117	Ph	Cl	CO*cy*Pr	N	–	6.631	6.704	6.755	6.923	6.879
118	2-Pyrrolyl	H	CO*cy*Pr	CH	–	6.495	6.597	6.619	6.789	6.744
119	Ph	Br	CO-4-Piperidine-N-Me	N	–	6.417	6.518	6.540	6.974	6.924
120	Ph	H	CO*cy*Pr	N	–	6.372	6.474	6.495	6.439	6.435
121	Ph-3-SO$_2$NH$_2$	H	CO*cy*Pr	CH	–	6.318	6.216	6.442	6.443	6.532
122	3-F-Ph	H	CO*cy*Pr	CH	–	6.082	6.015	6.158	6.151	6.224
123	6-Quinolyl	H	CO*cy*Pr	CH	–	6.000	5.983	6.124	6.394	6.353
124	Ph-4-SO$_2$NH$_2$	H	CO*cy*Pr	CH	–	6.000	6.027	6.124	6.311	6.417
125	2-F-Ph	H	CO*cy*Pr	CH	–	6.000	5.937	6.124	6.258	6.251
126	4-F-Ph	H	CO*cy*Pr	CH	–	6.000	6.102	6.124	6.230	6.272
127	2-MeO-Ph	H	CO*cy*Pr	N	–	5.798	5.900	5.922	6.422	6.427
128	4-MeO-Ph	H	CO*cy*Pr	N	–	5.301	5.403	5.425	5.918	6.005
129[b]	3-NO$_2$	4-Cl-3-CO$_2$H	H	–	–	7.585	7.523	7.059	7.142	7.265
130[b]	2-NO$_2$	3,5,-di-Cl-4-OH	H	–	–	7.284	7.373	7.032	7.103	7.298
131[b]	4-OMe	3,5,-di-Cl-4-OH	H	–	–	7.081	7.131	6.990	6.966	7.060
132[b]	3-Cl	3-Cl-4-OH	H	–	–	7.027	6.417	6.748	6.781	6.796
133[b]	4-Cl	4-Cl-3-CO$_2$H	H	–	–	6.963	7.314	6.973	6.933	7.021
134[b]	2-OMe	3-Cl-4-OH	H	–	–	6.857	7.074	6.888	6.844	6.983
135[b]	3-NO$_2$	H	H	–	–	6.851	6.635	6.588	6.428	6.441
136[b]	3-OMe	4-SMe	H	–	–	6.693	6.613	6.749	6.682	6.588
137[b]	2-OMe	3-OH	H	–	–	6.587	6.830	6.699	6.479	6.580
138[b]	H	3-Cl	H	–	–	6.521	6.370	6.421	6.379	6.311
139[b]	H	4-SMe	H	–	–	6.394	6.522	6.677	6.398	6.395
140[b]	4-Cl	3-OH	H	–	–	6.390	5.952	6.291	6.331	6.327
141[b]	H	H	Me	–	–	5.583	5.283	6.464	6.234	6.115
142[b]	3-NO$_2$	–	–	–	–	6.793	6.651	6.730	6.573	6.714
143[b]	3-Cl	–	–	–	–	6.337	5.913	6.349	6.432	6.315
144[b]	4-OMe	–	–	–	–	6.159	6.375	6.712	6.396	6.577

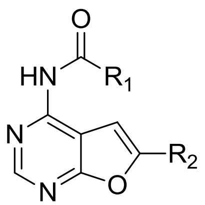
145[b]	Ph	H	CO$(CH_2)_4$Piperazinyl-N-Et	N	N	8.301	8.767	7.251	7.512	7.532
146[b]	Ph	H	CO-4-Piperidine-N-Me	N	N	8.046	7.802	7.232	7.881	7.909
147[b]	2-F-Ph	H	CO*n*Pr	CH	N	7.745	7.431	7.775	7.468	7.492
148[b]	Ph	H	CO*n*Pr	CH	N	7.252	7.310	7.377	7.286	7.254
149[b]	1-Naphthyl	H	CO*n*Pr	CH	N	6.618	6.922	6.901	7.206	7.087
150[b]	H	Ph	H	CH	H	5.301	6.252	6.913	6.018	6.155
151[b]	3-Cl-4-HO-Ph	H	CO*cy*Pr	N	–	8.155	8.753	7.468	7.462	7.441
152[b]	4-HO-Ph	H	CO-4-Piperidine-N-Me	N	–	7.921	8.455	6.969	7.924	7.902
153[b]	2-Thiazolyl	Br	CO*cy*Pent	N	–	7.796	8.190	7.241	7.782	7.735
154[b]	2-HO-Ph	H	CO*cy*Pr	N	–	7.444	7.952	6.910	7.269	7.077
155[b]	2-Furanyl	H	CO*cy*Pr	N	–	6.851	6.806	7.121	7.058	7.038
156[b]	Ph	Ph	CO*cy*Pr	N	–	6.382	6.371	7.107	7.223	7.107
157[b]	Ph	H	CO*cy*Pr	CH	–	6.303	5.836	6.394	6.439	6.237

[a] Ac = Acetyl, Br = Bromo, Bu = Butyl, Bn = Benzyl, Cl = Chloro, Et = Ethyl, Me = Methyl, Pent = Pentyl, Ph = Phenyl, Pr = Propyl, *n* = *normal*, *cy* = *cyclo*, *i* = *iso*, (±) = racemic mixture; [b] Test set.

Appendix: D. Structures and bioactivities of the compounds in Group III.



1-59
76-86[b]



60-75
87-91[b]

No.	R_1[a]	R_2[a]	R_3[a]	R_4[a]	R_5[a]	pIC_{50}				
						Expt.	Model			
							I	II	III	IV
1	H	H	7-Benzimidazole	4-CONH$(CH_2)_2$$NMe_2$-Ph	H	8.8	8.740	8.532	8.440	8.355
2	H	H	3-OMe-Ph	4-NH_2-Ph	H	8.6	8.525	8.332	8.189	8.135
3	H	H	3-OMe-Ph	4-SO_2Me-Ph	H	8.6	8.540	8.332	7.994	8.039
4	H	H	4-Pyridyl	4-Pyridyl	H	8.5	8.440	8.232	7.898	7.780
5	H	H	3-OMe-Ph	4-CH_2NHEt-Ph	H	8.4	8.340	8.132	8.202	8.112
6	H	H	7-Benzimidazole	4-Pyridyl	H	8.4	8.460	8.132	7.581	7.612
7	H	H	3-OMe-Ph	4-$CH_2$$NMe_2$-Ph	H	8.3	8.326	8.032	8.217	8.201
8	H	H	3-OMe-Ph	4-NHC(O)$(CH_2)_2$$NMe_2$-Ph	H	8.3	8.240	8.032	8.134	8.205
9	H	H	7-Benzimidazole	4-SO_2Me-Ph	H	8.3	8.240	8.032	7.669	7.823
10	H	H	7-Benzimidazole	4-COOH-Ph	H	8.3	8.240	8.032	7.775	7.783
11	H	H	3-OMe-Ph	4-Pyridyl	H	8.2	8.260	7.932	7.846	7.853
12	H	H	7-Benzimidazole	4-F-Ph	H	8.2	8.140	7.932	7.598	7.628
13	H	H	3-OMe-Ph	3-F-Ph	H	8.1	8.122	7.832	7.849	7.817
14	H	H	3-OMe-Ph	4-CONH$(CH_2)_2$$SO_2$Me-Ph	H	8.1	8.040	7.832	7.969	7.702
15	H	H	3-OMe-Ph	4-Pyridyl	Me	8.0	7.940	7.732	7.715	7.509
16	H	H	3-OMe-Ph	4-Pyridyl	*n*Pr	8.0	7.940	7.732	7.843	7.606
17	H	H	3-OMe-Ph	3-NH_2-Ph	H	8.0	8.060	7.732	7.952	7.850
18	H	H	3-OMe-Ph	4-SO_2NH$(CH_2)_2$$NMe_2$-Ph	H	8.0	7.989	7.732	7.750	7.919
19	H	H	3-OMe-Ph	4-OH-Ph	H	7.9	7.960	7.632	7.640	7.684
20	H	H	3-OMe-Ph	4-NHAc-Ph	H	7.9	7.880	7.632	7.995	7.811

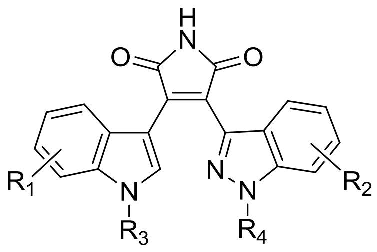
21	H	H	3-OMe-Ph	3,4-diF-Ph	H	7.8	7.740	7.532	7.411	7.516
22	H	H	3-OMe-Ph	4-NHSO$_2$Me-Ph	H	7.8	7.860	7.532	7.883	7.835
23	H	H	3-OMe-Ph	4-CN-Ph	H	7.8	7.740	7.532	7.843	7.822
24	H	H	3-OMe-Ph	4-CONH(CH$_2$)$_2$NMe$_2$-Ph	H	7.8	7.860	7.532	7.576	7.829
25	H	H	Ph	4-CONH(CH$_2$)$_2$NMe$_2$-Ph	H	7.7	7.760	7.432	7.696	7.547
26	H	H	3-OMe-Ph	3-OMe-Ph	H	7.6	7.660	7.332	7.683	7.462
27	H	H	3-Pyridyl	4-Pyridyl	H	7.5	7.560	7.232	7.520	7.271
28	H	H	3-OMe-Ph	4-OMe-Ph	H	7.5	7.440	7.232	7.113	7.154
29	H	H	3-OMe-Ph	4-Pyridyl	*i*Pr	7.5	7.440	7.232	7.134	7.131
30	H	Me	3-OMe-Ph	4-Pyridyl	H	7.2	7.140	6.932	7.311	7.004
31	H	H	Ph	4-Pyridyl	H	7.0	6.940	6.785	6.579	6.749
32	H	H	3-OEt-Ph	4-Pyridyl	H	7.0	7.060	6.785	7.428	7.195
33	H	H	Ph	3-OMe-4-OH-Ph	H	7.0	7.060	6.785	6.334	6.520
34	H	H	3-OMe-Ph	2-F-Ph	H	6.9	6.840	6.785	6.241	6.543
35	H	H	2-Me-Ph	4-Pyridyl	H	6.8	6.740	6.785	6.481	6.533
36	H	H	3-OMe-Ph	4-OAllyl-Ph	H	6.8	6.740	6.785	6.458	6.557
37	H	H	3-Br-Ph	4-Pyridyl	H	6.6	6.540	6.785	6.507	6.541
38	H	H	3-OCF$_3$-Ph	4-Pyridyl	H	6.5	6.560	6.768	6.167	6.247
39	H	H	3-NHCO*n*Pr-Ph	4-Pyridyl	H	6.2	6.260	6.468	6.069	6.194
40	H	H	Ph	4-F-Ph	H	6.2	6.260	6.468	6.312	6.340
41	H	H	4-OMe-Ph	4-Pyridyl	H	6.0	5.940	6.268	6.148	6.038
42	H	H	2-Pyridyl	4-Pyridyl	H	5.9	5.959	6.168	6.394	6.458
43	H	H	3-NHCH$_2$*cy*Pr-Ph	4-Pyridyl	H	5.7	5.640	5.968	5.904	5.847
44	H	H	2-OMe-Ph	4-Pyridyl	H	5.6	5.660	5.868	5.812	6.054
45	H	H	*i*Pr	3-OMe-4-OH-Ph	H	5.6	5.660	5.868	6.127	5.986
46	H	H	3-OMe-Ph	4-O(4-F)Bn-Ph	H	5.6	5.660	5.868	5.582	5.717
47	H	H	3-OMe-Ph	4-NEt$_2$-Ph	H	5.6	5.660	5.868	5.952	6.015
48	H	H	H	3-OMe-4-OH-Ph	H	5.5	5.560	5.768	5.827	6.104
49	H	H	3-NH*n*Pr-Ph	4-Pyridyl	H	5.4	5.340	5.668	5.883	5.875
50	H	Me	3-OMe-Ph	4-Pyridyl	Me	5.2	5.260	5.468	6.376	6.143
51	Cl	H	Ph	3-OMe-4-OH-Ph	H	5.0	5.060	5.268	5.382	5.436
52	Ph	H	Ph	3-OMe-4-OH-Ph	H	5.0	5.060	5.268	5.473	5.795
53	H	Me	Ph	3-OMe-4-OH-Ph	H	5.0	5.060	5.268	5.129	5.402

54	H	H	Bn	3-OMe-4-OH-Ph	H	5.0	5.060	5.268	5.367	5.362
55	H	H	3-OMe-Ph	2-OMe-Ph	H	4.6	4.660	4.868	5.278	5.191
56	H	H	4-Me-Ph	4-Pyridyl	H	4.5	4.560	4.768	5.039	5.299
57	H	H	3-NO$_2$-Ph	4-Pyridyl	H	4.5	4.560	4.768	5.097	5.110
58	H	H	3-OMe-Ph	4-Me-Ph	H	4.5	4.560	4.768	5.982	5.811
59	H	H	3-OMe-Ph	4-Ph-Ph	H	4.5	4.560	4.768	5.050	5.172
60	*cy*Pent	3-Pyridyl	–	–	–	8.3	8.240	8.032	8.029	7.797
61	*cy*Pent	Ph	–	–	–	7.6	7.540	7.332	7.137	7.278
62	*cy*Pent	4-OMe-Ph	–	–	–	7.5	7.440	7.232	6.931	7.142
63	CHMe$_2$	4-OMe-Ph	–	–	–	6.9	6.840	6.785	6.773	6.758
64	*cy*Pr	Ph	–	–	–	6.8	6.740	6.785	6.585	6.935
65	*cy*Pent	4-Me-Ph	–	–	–	6.6	6.660	6.785	6.712	6.600
66	*cy*Pr	4-OMe-Ph	–	–	–	6.5	6.560	6.768	6.672	6.592
67	(CH$_2$)$_4$Me	4-OMe-Ph	–	–	–	6.4	6.340	6.668	6.185	6.413
68	(CH$_2$)$_2$SMe	4-OMe-Ph	–	–	–	6.3	6.360	6.568	6.332	6.372
69	*cy*Pent	4-Cl-Ph	–	–	–	6.3	6.360	6.568	6.358	6.363
70	*cy*Pr	4-Me-Ph	–	–	–	6.2	6.260	6.468	6.352	6.474
71	*cy*Pr	4-Cl-Ph	–	–	–	6.0	6.060	6.268	6.117	6.248
72	*cy*He	4-OMe-Ph	–	–	–	5.7	5.760	5.968	5.958	5.857
73	Pyrrolidinyl	4-OMe-Ph	–	–	–	5.7	5.760	5.968	6.005	6.074
74	2-Furanyl	4-OMe-Ph	–	–	–	5.6	5.660	5.868	5.831	5.953
75	3-F-Ph	4-OMe-Ph	–	–	–	4.5	4.560	4.768	4.932	5.086
76^b	H	H	3-OMe-Ph	3-SO$_2$Me-Ph	H	8.4	7.570	6.785	5.921	6.353
77^b	H	H	3-OMe-Ph	4-O(CH$_2$)$_2$NMe$_2$-Ph	H	8.2	7.781	6.785	6.574	6.507
78^b	H	H	3-OMe-Ph	4-F-Ph	H	8.1	7.329	6.785	7.157	7.334
79^b	H	H	3-OMe-Ph	Ph	H	8.0	7.782	6.785	7.371	6.851
80^b	H	H	3-OMe-Ph	4-Pyridyl	Et	7.9	7.188	6.785	7.786	7.299
81^b	H	H	3-OMe-Ph	4-COOH-Ph	H	7.5	7.787	6.785	7.024	7.440
82^b	H	H	3-OMe-2-Pyridyl	4-Pyridyl	H	7.4	6.945	6.785	7.767	7.480
83^b	H	H	2-Thiazolyl	4-Pyridyl	H	7.4	6.258	6.785	6.648	7.091
84^b	H	H	3-Me-Ph	4-Pyridyl	H	6.8	6.472	6.785	6.225	5.986
85^b	H	H	3-NHAc-Ph	4-Pyridyl	H	6.8	7.396	6.785	6.461	6.541
86^b	H	H	3-F-Ph	4-Pyridyl	H	6.5	7.402	6.785	7.619	7.439

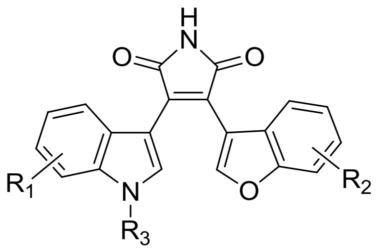
87[b]	*cy*Pr	3-Pyridyl	–	–	–	8.3	7.394	6.785	7.949	7.673
88[b]	*cy*Pent	4-F-Ph	–	–	–	6.7	6.786	6.785	6.534	6.763
89[b]	CH_2*cy*Pent	4-OMe-Ph	–	–	–	6.5	6.592	6.785	6.184	6.274
90[b]	*cy*Pr	4-F-Ph	–	–	–	6.2	6.313	6.785	6.554	6.650
91[b]	Morpholinyl	4-OMe-Ph	–	–	–	4.5	5.545	6.785	6.467	6.038

[a] Ac = Acetyl, Br = Bromo, Bu = Butyl, Bn = Benzyl, Cl = Chloro, Et = Ethyl, He = Hexyl, Me = Methyl, Pent = Pentyl, Ph = Phenyl, Pr = Propyl, *n* = *normal*, *cy* = *cyclo*; [b] Test set.

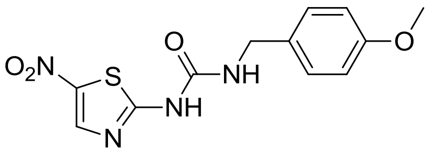
Appendix: E. Structures and bioactivities of the compounds in Group IV.



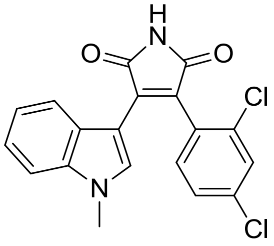
1-25
67-71[b]



26-64
72-85[b]



65



66

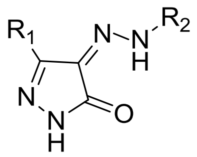
No.	R_1[a]	R_2[a]	R_3[a]	R_4[a]	pIC_{50}				
					Expt.	Model			
						I	II	III	IV
1	5-Br	H	Me	H	8.456	7.477	8.408	7.818	7.736
2	5-F	H	H	H	7.943	6.648	7.895	7.543	7.415
3	5-F	H	$(CH_2)_2OMe$	H	7.444	7.426	7.396	7.145	7.127
4	5-F	H	Me	H	7.310	7.292	7.358	7.271	7.150
5	5-NO_2	H	$(CH_2)_2OMe$	H	7.284	7.302	7.236	7.295	7.255
6	5-CN	H	H	H	7.092	7.073	7.043	6.892	6.921
7	5-OBn	H	$(CH_2)_3OH$	H	7.000	6.982	6.952	6.757	6.865
8	H	H	Me	Me	6.921	6.903	6.873	6.495	6.768
9	5-OBn	H	$(CH_2)_2OMe$	H	6.886	6.868	6.838	6.787	6.820
10	5-F	5-Cl	$(CH_2)_2OMe$	H	6.783	6.801	6.802	6.665	6.703
11	5-Cl	H	$(CH_2)_2OMe$	H	6.420	6.402	6.372	6.540	6.464
12	6-F	H	H	H	6.272	6.253	6.320	6.486	6.570
13	5-OBn	5-Cl	$(CH_2)_2OMe$	H	6.125	6.143	6.173	6.243	6.260
14	5-Br	H	H	H	6.071	6.088	6.119	6.376	6.400
15	5-Cl	H	Me	H	5.959	6.263	5.910	6.141	6.174
16	5-OBn	H	Me	H	5.921	5.939	5.969	6.025	6.133
17	6-NO_2	H	Me	H	5.886	5.904	5.934	5.987	6.212
18	H	H	$(CH_2)_3OH$	H	5.745	5.727	5.793	5.684	5.967
19	H	H	$(CH_2)_2OMe$	H	5.602	6.122	5.650	6.109	6.071
20	H	H	Me	H	5.585	5.573	5.633	5.680	5.817

21	6-Cl	H	Me	H	5.509	5.347	5.556	5.673	5.735
22	H	5-Cl	Me	H	5.509	5.527	5.556	5.728	5.816
23	H	5-Cl	$(CH_2)_3OH$	H	5.444	5.462	5.492	5.555	5.647
24	5-Cl	H	$(CH_2)_3OH$	H	5.301	5.283	5.349	5.787	5.791
25	H	H	H	H	4.921	4.939	4.969	5.411	5.615
26	7-CH_2OMe	H	Me	–	9.638	8.838	9.591	8.718	8.612
27	5-F	6-CH_2OH	Me	–	9.456	9.438	9.408	8.590	8.572
28	7-CH_2OMe	6-CH_2OH	Me	–	9.137	9.119	9.089	8.710	8.667
29	5-F-6-Cl	6-CH_2OH	Me	–	9.022	9.004	8.974	8.289	8.037
30	5-F	6-OH	Me	–	8.456	8.438	8.408	7.954	7.928
31	7-CH_2OH	H	Me	–	8.268	8.250	8.220	8.223	8.132
32	5-Br	H	Me	–	8.155	8.137	8.107	7.766	7.752
33	5-C≡CH	H	Me	–	8.018	8.036	7.970	7.586	7.580
34	7-$(CH_2)_2COOEt$	H	Me	–	7.991	8.009	7.943	8.072	7.819
35	5-CN	6-CH_2OH	Me	–	7.879	7.897	7.832	7.807	7.759
36	6-OH	5-F	Me	–	7.854	7.872	7.831	7.701	7.738
37	6-OH	H	Me	–	7.824	7.806	7.776	7.754	7.628
38	5-C≡C*cy*Pr	5-F	Me	–	7.793	7.811	7.745	7.571	7.496
39	5-F	6-CH_2OMe	Me	–	7.623	8.746	7.672	7.848	7.676
40	5-F	H	Me	–	7.585	7.567	7.537	7.570	7.405
41	5-Cl	5-F	Me	–	7.377	7.359	7.329	7.211	7.258
42	5-Br	6-$OCH_2CH=CH_2$	Me	–	7.316	7.298	7.268	7.228	7.194
43	7-OH	H	Me	–	7.260	7.242	7.212	7.269	7.206
44	5,7-diBr	7-OMe	Me	–	7.052	7.070	7.004	7.274	7.157
45	5-OMe	H	Me	–	6.903	6.885	6.951	6.826	6.959
46	6-OBn	5-F	Me	–	6.796	6.778	6.773	6.692	6.722
47	H	7-OMe	Me	–	6.745	6.763	6.793	7.058	6.963
48	5-F-6-Cl	H	Me	–	6.735	6.753	6.783	6.865	6.802
49	7-OBn	H	Me	–	6.658	6.639	6.638	6.650	6.626
50	5-F-6-I	7-OMe	Me	–	6.607	6.590	6.655	6.641	6.780
51	5-F-6-Cl	7-OMe	Me	–	6.585	6.567	6.633	6.586	6.627
52	5-Br	6-O-(4-OMe)-Bn	Me	–	6.475	6.493	6.523	6.644	6.601
53	5-F	H	H	–	6.444	6.448	6.492	6.715	6.524

54	5-OMe-6-Cl	H	Me	–	6.357	6.339	6.405	6.445	6.496
55	5-OBn	H	Me	–	6.301	6.319	6.253	6.215	6.325
56	H	5-Br	Me	–	6.260	6.671	6.308	6.689	6.789
57	H	5-F	H	–	6.174	6.192	6.222	6.402	6.412
58	5-OH	H	Me	–	6.161	7.051	6.209	6.479	6.600
59	5,6-methylenedioxy	5-F	Me	–	6.149	6.167	6.197	6.634	6.553
60	6-CF$_3$	7-OMe	Me	–	6.081	6.099	6.129	6.239	6.569
61	5-F-6-Cl	6-OMe	Me	–	6.060	6.079	6.109	6.434	6.492
62	5-OBn	H	H	–	5.783	5.800	5.830	6.026	6.035
63	5-F-6-Cl	6-OCH$_2$*cy*Bu	Me	–	5.388	5.407	5.436	6.080	6.008
64	5-F-6-(4-Cl)-Ph	7-OMe	Me	–	5.145	5.163	5.193	5.656	5.736
65	–	–	–	–	7.379	7.361	7.331	7.350	7.227
66	–	–	–	–	7.301	7.283	7.253	7.165	7.109
67[b]	H	H	Me	(CH$_2$)$_2$OMe	7.310	7.488	6.445	6.395	6.724
68[b]	7-OBn	H	(CH$_2$)$_2$OMe	H	6.398	6.495	6.789	7.059	6.748
69[b]	6-OBn	H	(CH$_2$)$_2$OMe	H	6.347	6.525	6.768	6.939	6.775
70[b]	5-NO$_2$	H	Me	H	6.337	7.122	6.529	6.725	6.705
71[b]	6-Me	5-Cl	(CH$_2$)$_2$OMe	H	5.620	5.974	6.515	6.250	6.194
72[b]	5-Br	6-CH$_2$OH	Me	–	9.292	9.547	7.234	7.915	7.893
73[b]	7-(CH$_2$)$_2$COOH	H	Me	–	8.921	8.541	7.313	8.285	8.040
74[b]	7-CH$_2$OH	6-CH$_2$OH	Me	–	8.292	9.003	7.939	7.899	8.113
75[b]	5-Br	7-OMe	Me	–	8.125	8.735	7.118	7.353	7.190
76[b]	5-Br	6-OCH$_2$C≡CH	Me	–	7.597	7.192	7.054	7.325	7.144
77[b]	5-I	H	Me	–	7.462	6.346	6.860	6.863	7.197
78[b]	H	H	Me	–	7.456	5.717	6.427	6.803	6.793
79[b]	5-I	5-F	Me	–	6.745	6.715	6.829	7.056	7.228
80[b]	5-OBn	H	(CH$_2$)$_3$OH	–	6.658	7.062	6.801	6.904	6.889
81[b]	5-OMe-6-I	H	Me	–	6.652	5.873	6.813	6.561	6.765
82[b]	5-*cy*Pr	H	Me	–	6.629	7.712	6.830	6.902	7.086
83[b]	1*H*-benzo[g]	5,6-diF	Me	–	6.503	6.902	6.816	7.242	7.061
84[b]	6-OBn	H	Me	–	6.046	6.503	6.562	6.516	6.621
85[b]	5-F-6-Cl	6-OCH$_2$*cy*Pr	Me	–	5.889	5.429	6.354	6.604	6.544

[a] Br = Bromo, Bu = Butyl, Bn = Benzyl, Cl = Chloro, Et = Ethyl, Me = Methyl, Ph = Phenyl, Pr = Propyl, *cy* = *cyclo* ; [b] Test set.

Appendix: F. Structures and bioactivities of the compounds in Group V.



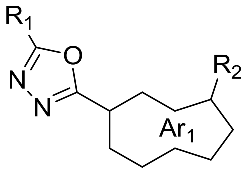
1-62

No.	R_1[a]	R_2[a]	pIC_{50}				
			Expt.	Model			
				I	II	III	IV
1	3,4-(OMe)$_2$-Ph	3-Cl-Ph	9.398	9.286	9.374	9.025	9.043
2	3,4-(OMe)$_2$-Ph	3-OMe-Ph	9.398	9.286	9.374	9.112	9.075
3	4-OMe-Ph	3-OMe-Ph	9.097	8.985	9.073	8.864	8.840
4	4-OMe-Ph	3-Cl-Ph	8.721	8.833	8.697	8.521	8.574
5	4-OMe-Ph	4-OMe-Ph	8.699	8.329	8.675	8.366	8.384
6	4-OMe-3-Pyridyl	4-OMe-Ph	8.699	8.705	8.675	8.687	8.650
7	4-OMe-Ph	4-Cl-Ph	8.699	8.587	8.675	8.547	8.512
8	4-OMe-Ph	3-Pyridyl	8.699	8.587	8.675	8.684	8.601
9	3,4-(OMe)$_2$-Ph	4-Cl-Ph	8.699	8.587	8.675	8.535	8.472
10	3,4-(OMe)$_2$-Ph	4-CN-Ph	8.699	8.750	8.675	8.544	8.551
11	3,4-(OMe)$_2$-Ph	4-NEt$_2$-Ph	8.699	8.587	8.675	8.514	8.551
12	3,4-(OMe)$_2$-Ph	4-Morpholinyl-Ph	8.523	8.411	8.499	8.488	8.436
13	3,4-(OMe)$_2$-Ph	Ph	8.456	8.568	8.432	8.486	8.437
14	3,4-(OMe)$_2$-Ph	4-Pyridyl	8.456	8.568	8.432	8.328	8.337
15	4-OMe-3-Pyridyl	Ph	8.347	8.235	8.323	8.183	8.170
16	3-OMe-Ph	3-OMe-Ph	8.301	8.405	8.325	8.372	8.331
17	3-OMe-Ph	4-COOH-Ph	8.187	8.076	8.163	8.166	8.165
18	4-OMe-3-Pyridyl	4-COOH-Ph	8.114	8.226	8.090	8.269	8.244
19	3-OMe-Ph	3-Cl-Ph	8.097	8.209	8.072	8.127	8.077
20	3-OMe-Ph	2-Pyridyl	8.097	7.988	8.073	7.976	8.015
21	4-OMe-Ph	4-Morpholinyl-Ph	8.097	7.985	8.073	7.887	7.947
22	3-OMe-Ph	4-Cl-Ph	8.046	8.158	8.022	7.968	8.012
23	3-OMe-Ph	4-NMe$_2$-Ph	8.046	7.934	8.022	8.040	7.945
24	4-OMe-Ph	4-Pyridyl	7.854	8.199	7.878	7.839	7.875
25	4-OMe-3-Pyridyl	4-Pyridyl	7.796	7.908	7.772	7.695	7.715
26	3-OMe-Ph	4-OMe-Ph	7.745	7.857	7.768	7.999	7.977
27	Ph	4-OMe-Ph	7.699	7.587	7.675	7.702	7.695
28	3-Pyridyl	4-OMe-Ph	7.699	7.811	7.675	7.673	7.714
29	Ph	4-Cl-Ph	7.638	7.526	7.614	7.264	7.322
30	3-OMe-Ph	4-Pyridyl	7.638	7.527	7.662	7.507	7.458
31	4-OMe-3-Pyridyl	3-Pyridyl	7.638	7.750	7.662	7.828	7.815
32	3-OMe-Ph	Ph	7.409	7.516	7.404	7.571	7.537
33	3,4-(OMe)$_2$-Ph	2-Cl-Ph	7.357	7.363	7.381	7.585	7.761
34	4-OMe-3-Pyridyl	4-Morpholinyl-Ph	7.301	7.413	7.325	7.635	7.628
35	3-OMe-Ph	4-Morpholinyl-Ph	7.229	7.341	7.254	7.512	7.546
36	3,4-(OMe)$_2$-Ph	4-NMe$_2$-Ph	7.060	7.173	7.085	7.645	7.788
37	Ph	Ph	7.004	7.116	7.028	6.784	6.777

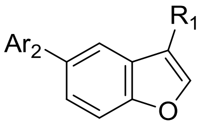
38	4-OMe-Ph	4-NMe$_2$-Ph	6.959	7.070	6.983	7.360	7.551
39	Ph	3-Cl-Ph	6.921	7.033	6.944	6.925	6.972
40	3-Pyridyl	2-Cl-Ph	6.699	6.586	6.675	6.483	6.558
41	4-OMe-3-Pyridyl	2-OMe-Ph	6.569	6.615	6.592	6.875	6.974
42	Me	2-Cl-Ph	6.337	6.439	6.361	6.058	6.159
43	2-Pyridyl	4-OMe-Ph	6.194	6.306	6.218	6.695	6.755
44	2-Pyridyl	2-Cl-Ph	6.187	6.142	6.211	6.263	6.277
45	Ph	2-Cl-Ph	6.143	6.255	6.167	6.208	6.319
46	3-OMe-Ph	2-Cl-Ph	6.071	6.183	6.095	6.497	6.529
47	H	2-Cl-Ph	5.827	5.715	5.851	5.874	5.994
48	*i*Pr	2-Cl-Ph	5.432	5.544	5.456	5.826	5.815
49	Ph	2-OMe-Ph	5.398	5.510	5.422	5.916	5.949
50[b]	3,4-(OMe)$_2$-Ph	4-OMe-Ph	8.699	9.291	8.636	8.740	8.730
51[b]	4-OMe-Ph	4-CN-Ph	8.699	8.460	8.464	8.523	8.407
52[b]	4-OMe-3-Pyridyl	4-Cl-Ph	8.523	8.541	8.043	7.893	7.853
53[b]	3,4-(OMe)$_2$-Ph	3-Pyridyl	8.456	8.956	8.057	8.426	8.334
54[b]	4-OMe-3-Pyridyl	3-OMe-Ph	8.432	9.106	8.259	8.546	8.431
55[b]	3-OMe-Ph	4-CN-Ph	8.398	8.024	7.921	7.977	7.999
56[b]	4-OMe-3-Pyridyl	4-CN-Ph	8.081	8.463	7.982	7.661	7.710
57[b]	3,4,5-(OMe)$_3$-Ph	4-OMe-Ph	8.046	8.043	7.607	8.798	8.625
58[b]	4-OMe-Ph	4-NEt$_2$-Ph	7.796	7.452	7.571	7.374	7.521
59[b]	4-Pyridyl	4-OMe-Ph	7.523	7.205	7.769	7.249	7.321
60[b]	3-OMe-Ph	3-Pyridyl	7.523	7.979	7.742	7.751	7.707
61[b]	Ph	3-OMe-Ph	7.456	7.487	7.689	7.619	7.684
62[b]	4-Pyridyl	2-Cl-Ph	6.602	6.732	6.761	6.415	6.425

[a] Cl = Chloro, Et = Ethyl, Me = Methyl, Ph = Phenyl, *i* = *iso*; [b] Test set.

Appendix: G. Structures and bioactivities of the compounds in Group VI.

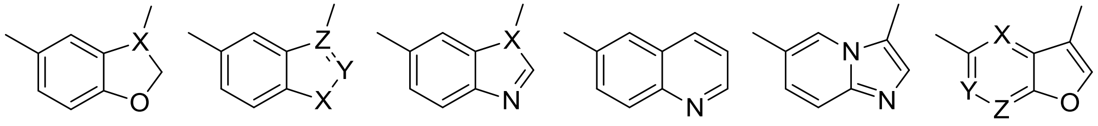


1-45
50-60[b]

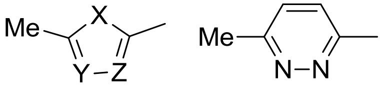


46-49
61[b]

Ar$_1$ =



Ar$_2$ =



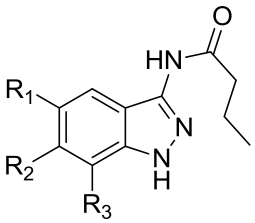
No.	R$_1$[a]	R$_2$[a]	Ar$_1$	Ar$_2$	X	Y	Z	pIC$_{50}$				
								Expt.	Model			
									I	II	III	IV
1	-S-CH$_2$-(3-CN)-Ph	4-OMe-Ph	r	–	N	–	–	8.638	8.558	8.566	8.378	8.429
2	-S-CH$_2$-(3-CF$_3$)-Ph	4-OMe-Ph	r	–	N	–	–	8.602	8.522	8.530	8.408	8.426
3	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	–	r	–	S	–	–	8.509	8.429	8.437	8.265	8.314
4	-S-CH$_2$-(3-CN)-Ph	4-OMe-Ph	q	–	O	CH	C	8.456	8.504	8.384	8.397	8.336
5	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	H	q	–	O	CH	C	8.310	8.390	8.238	8.171	8.207
6	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	H	t	–	–	–	–	8.187	8.107	8.115	7.987	7.984
7	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	4-OMe-Ph	r	–	N	–	–	8.066	7.986	7.994	8.031	8.087
8	-S-CH$_2$-(3-CF$_3$)-Ph	Me	r	–	N	–	–	8.027	7.947	7.955	7.721	7.877
9	-S-CH$_2$-(3-Cl-4-OMe)-Ph	H	p	–	CH	–	–	7.886	7.806	7.814	7.778	7.521
10	NH$_2$	4-S(O)Me-Ph	q	–	O	CH	C	7.886	7.806	7.814	7.257	7.446
11	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	–	s	–	–	–	–	7.745	7.664	7.672	7.970	7.850
12	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	4-OMe-Ph	q	–	O	CH	C	7.602	7.682	7.530	7.990	7.854
13	-S-CH$_2$-(3-CN-4-OMe)-Ph	H	p	–	CH	–	–	7.553	7.473	7.481	7.877	7.470
14	Me	4-S(O)Me-Ph	q	–	O	CH	C	7.456	7.376	7.384	7.030	7.094
15	Me	4-S(O)Et-Ph	q	–	O	CH	C	7.456	7.376	7.384	7.106	7.198
16	Me	4-S(O)$_2$Me-Ph	q	–	O	CH	C	7.377	7.296	7.305	6.882	7.226
17	-S-CH$_2$-(3-F-4-OMe)-Ph	H	p	–	CH	–	–	7.357	7.437	7.285	7.442	7.185
18	Me	4-CO$_2$H-Ph	q	–	O	CH	C	7.301	7.221	7.229	6.881	7.146
19	-S-CH$_2$-(3-F)-Ph	H	p	–	CH	–	–	7.268	7.187	7.196	6.761	6.953
20	Me	4-OMe-Ph	q	–	O	CH	C	7.268	7.187	7.195	7.091	7.083

21	-S-CH$_2$-(3-F-4-OMe)-Ph	–	p	–	O	–	–	7.187	7.267	7.169	7.054	7.128
22	Me	4-SMe-Ph	q	–	O	CH	C	7.18	7.26	7.108	6.689	7.054
23	-S-CH$_2$-(3-CF$_3$)-Ph	H	p	–	CH	–	–	7.167	7.247	7.239	6.980	7.209
24	Me	4-CONH$_2$-Ph	q	–	O	CH	C	7.161	7.081	7.092	6.909	7.100
25	Me	3-OMe-Ph	q	–	O	CH	C	7.131	7.211	7.141	6.862	7.025
26	SH	4-S(O)Me-Ph	q	–	O	CH	C	7.027	7.107	7.067	6.980	6.935
27	Me	4-F-Ph	q	–	O	CH	C	6.959	7.039	7.031	6.866	6.929
28	OH	4-S(O)Me-Ph	q	–	O	CH	C	6.886	6.966	6.958	6.976	6.953
29	Me	4-C(OH)Me-Ph	q	–	O	CH	C	6.796	6.876	6.868	6.933	6.858
30	Me	4-OH-Ph	q	–	O	CH	C	6.745	6.825	6.817	6.801	6.808
31	-S-CH$_2$-(2-Cl)-Ph	H	p	–	CH	–	–	6.699	6.779	6.771	6.551	6.708
32	Me	4-CO$_2$Me-Ph	q	–	O	CH	C	6.699	6.779	6.771	6.688	6.854
33	Me	4-S(O)Me-Ph	r	–	N	–	–	6.699	6.619	6.771	6.746	6.628
34	-S-CH$_2$-Ph	H	p	–	CH	–	–	6.678	6.597	6.750	6.760	6.689
35	-S-CH$_2$-(4-OMe)-Ph	H	p	–	CH	–	–	6.658	6.612	6.730	6.891	6.741
36	Me	4-S(O)Me-Ph	u	–	N	CH	CH	6.569	6.489	6.641	6.752	6.586
37	-S-CH$_2$-(4-Cl)-Ph	H	p	–	CH	–	–	6.553	6.472	6.625	6.517	6.647
38	Me	4-P(O)(OMe)$_2$-Ph	q	–	O	CH	C	6.553	6.633	6.625	6.712	6.867
39	-NH-CH$_2$-(3-F)-Ph	H	p	–	CH	–	–	6.481	6.402	6.418	6.523	6.537
40	-S-CH$_2$-(3-CO$_2$Me)-Ph	H	p	–	CH	–	–	6.201	6.281	6.273	6.560	6.638
41	-CH$_2$-S-(3-F)-Ph	H	p	–	CH	–	–	6.167	6.136	6.239	6.429	6.383
42	Me	4-S(O)Me-Ph	t	–	–	–	–	6.000	6.080	6.072	6.609	6.320
43	Me	4-S(O)Me-Ph	u	–	CH	CH	N	6.000	6.007	6.072	6.280	6.366
44	Me	4-S(O)Me-Ph	u	–	CH	N	CH	6.000	5.920	6.072	6.330	6.305
45	-O-CH$_2$-(3-F)-Ph	H	p	–	CH	–	–	5.000	5.080	5.072	6.104	5.755
46	4-S(O)Me-Ph	–	–	v	S	N	N	7.143	7.063	7.070	6.728	6.947
47	4-S(O)Me-Ph	–	–	v	N	O	N	6.167	6.247	6.239	6.474	6.369
48	4-S(O)Me-Ph	–	–	w	–	–	–	6.000	6.080	6.072	6.364	6.386
49	4-S(O)Me-Ph	–	–	v	N	N	O	6.000	6.080	6.072	6.459	6.306
50^b	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	H	p	–	CH	–	–	8.244	8.116	7.329	7.826	7.702
51^b	-S-CH$_2$-(3-CF$_3$)-Ph	4-OMe-Ph	q	–	O	CH	C	8.076	8.538	7.497	8.362	8.167
52^b	-S-CH$_2$-(3-CF$_3$-4-OMe)-Ph	H	q	–	NH	N	C	7.796	8.010	7.368	7.430	7.655
53^b	Me	4-S(O)$_2$Et-Ph	q	–	O	CH	C	7.420	7.109	7.255	6.862	7.209

54[b]	Me	4-Pyridyl	q	–	O	CH	C	7.114	6.798	7.137	6.857	6.797
55[b]	-S-CH_2-(3-CN)-Ph	H	p	–	CH	–	–	7.041	6.763	7.058	6.621	6.874
56[b]	-S-CH_2-(3-Cl)-Ph	H	p	–	CH	–	–	7.027	7.010	6.963	6.591	6.706
57[b]	Me	4-COMe-Ph	q	–	O	CH	C	7.000	7.111	7.001	6.634	7.054
58[b]	Me	3-S(O)Me-Ph	q	–	O	CH	C	6.678	6.873	7.091	6.812	6.968
59[b]	-CH_2-CH_2-(3-F)-Ph	H	p	–	CH	–	–	6.469	6.268	6.748	6.419	6.714
60[b]	Me	4-S(O)Me-Ph	q	–	S	CH	C	6.102	6.482	7.114	6.759	6.850
61[b]	4-S(O)Me-Ph	–	–	v	NH	N	N	7.092	6.932	6.967	6.804	6.676

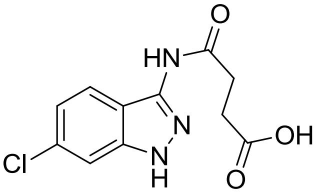
[a] Cl = Chloro, Et = Ethyl, Me = Methyl, Ph = Phenyl; [b] Test set.

Appendix: H. Structures and bioactivities of the compounds in Group VII.



1-45

47-58[b]



46

No.	R_1[a]	R_2[a]	R_3[a]	pIC$_{50}$				
				Expt.	Model			
					I	II	III	IV
1	Br	4-OH-Ph	H	8.301	8.241	8.285	8.015	7.839
2	Ph	F	F	8.222	8.162	8.206	7.782	7.760
3	Ph	Cl	H	8.155	8.095	8.139	7.938	7.731
4	4-NH_2-Ph	Cl	H	8.097	8.037	8.081	7.861	7.982
5	4-OH-Ph	Cl	H	8.000	8.060	7.984	7.942	7.948
6	Ph	Cl	Cl	7.921	8.060	7.904	7.895	7.835
7	4-Me-Ph	Cl	H	7.854	7.794	7.838	7.333	7.481
8	2-Furanyl	4-OH-Ph	H	7.854	7.794	7.838	7.878	7.769
9	4-NO_2-Ph	Cl	H	7.745	7.778	7.729	7.680	7.601
10	4-F-Ph	Cl	H	7.658	7.718	7.641	7.833	7.736
11	4-OBn-Ph	Cl	H	7.638	7.578	7.622	7.697	7.394
12	H	3,4-diOH-Ph	H	7.569	7.508	7.553	7.292	7.066
13	H	4-OH-Ph	H	7.357	7.296	7.341	6.843	6.695
14	4-Et-Ph	4-OH-Ph	H	7.310	7.370	7.294	6.614	6.849
15	4-Pyridyl	Cl	H	7.301	7.241	7.285	7.528	7.478
16	H	2-Cl-4-OH-Ph	H	7.260	7.320	7.244	7.352	7.240
17	Ph	4-OH-Ph	H	7.260	7.200	7.244	7.085	7.099
18	H	4-NH_2-Ph	H	7.119	7.179	7.103	6.673	6.517
19	H	3-OH-Ph	H	7.027	7.087	7.043	6.994	6.854
20	3-Furanyl	4-OH-Ph	H	7.009	7.069	6.992	6.901	6.996
21	3-Pyridyl	4-OH-Ph	H	6.827	6.767	6.811	6.299	6.689
22	H	CF_3	H	6.710	6.770	6.694	6.573	6.518
23	H	Br	H	6.703	6.643	6.687	6.571	6.533
24	H	3-Furanyl	H	6.585	6.525	6.569	6.483	6.343
25	H	4-Pyridyl	H	6.180	6.240	6.163	5.878	6.008
26	H	4-OBn-Ph	H	5.927	5.867	5.943	5.373	5.610
27	H	Ph	H	5.892	5.952	5.909	6.123	5.747
28	Ph	Ph	H	5.764	5.825	5.78	6.250	6.200
29	H	(*E*)-CH=CHMe	H	5.609	5.669	5.626	5.652	5.832
30	H	3,5-diF-Ph	H	5.539	5.498	5.554	5.217	5.636
31	H	4-$_{t}$Bu-Ph	H	5.419	5.479	5.435	5.152	5.302
32	H	4-F-Ph	H	5.300	5.346	5.316	5.555	5.635
33	H	3,5-diCl-Ph	H	5.267	5.207	5.282	5.014	5.598
34	H	CH_2Bn	H	5.069	5.009	5.085	5.521	5.439

35	4-OBn-Ph	4-OBn-Ph	H	5.000	5.060	5.016	5.086	5.572
36	H	4-CF$_3$-Ph	H	4.977	4.917	4.993	4.928	5.072
37	H	3-Pyridyl	H	4.963	4.903	4.979	5.605	5.376
38	H	4-NMe$_2$-Ph	H	4.912	4.852	4.928	4.828	4.979
39	H	2-OH-5-Pyridyl	H	4.775	4.835	4.791	5.812	5.537
40	H	4-Me-Ph	H	4.767	4.706	4.783	4.878	4.846
41	H	4-CN-Ph	H	4.688	4.747	4.704	5.270	5.111
42	H	4-NO$_2$-Ph	H	4.000	4.060	4.016	4.559	4.619
43	H	4-OMe-Ph	H	4.000	4.060	4.016	4.469	4.518
44	H	4-Et-Ph	H	4.000	4.060	4.016	4.741	4.466
45	H	4-Cl-Ph	H	4.000	4.137	4.016	4.458	4.571
46	–	–	–	5.928	5.989	5.944	5.734	6.005
47[b]	3-Furanyl	Cl	H	7.959	6.917	6.475	6.437	6.775
48[b]	4-Et-Ph	Cl	H	7.921	7.721	6.823	6.748	7.145
49[b]	Br	Cl	H	7.260	6.796	6.266	6.478	6.891
50[b]	4-OH-Ph	4-OH-Ph	H	7.252	6.862	6.665	7.271	7.120
51[b]	H	Cl	H	6.451	6.204	6.291	5.821	6.152
52[b]	H	3-Thienyl	H	6.029	6.926	6.186	7.334	6.334
53[b]	H	2-NH$_2$-5-Pyridyl	H	5.963	4.856	5.954	5.939	5.844
54[b]	H	2-Cl-Ph	H	5.869	6.404	6.209	6.169	6.499
55[b]	H	Bn	H	5.383	4.656	5.554	5.467	5.555
56[b]	H	4-OCF$_3$-Ph	H	5.260	4.684	6.192	5.618	5.933
57[b]	H	4-SMe-Ph	H	5.145	5.244	5.507	4.808	5.003
58[b]	H	3,4-methylenedioxy-Ph	H	4.601	5.228	6.170	6.836	6.223

[a] Br = Bromo, Bn = Benzyl, Cl = Chloro, Et = Ethyl, Me = Methyl, Ph = Phenyl; [b] Test set.

Appendix: I. Features selected for Group I by RBE-RF.

Name	Block	Description
J	Topological descriptors	Balaban distance connectivity index
MAXDP	Topological descriptors	maximal electrotopological positive variation
D/Dr05	Topological descriptors	distance/detour ring index of order 5
D/Dr09	Topological descriptors	distance/detour ring index of order 9
T(N..N)	Topological descriptors	sum of topological distances between N..N
SIC1	Information indices	structural information content (neighborhood symmetry of 1-order)
IC2	Information indices	information content index (neighborhood symmetry of 2-order)
SIC2	Information indices	structural information content (neighborhood symmetry of 2-order)
IC3	Information indices	information content index (neighborhood symmetry of 3-order)
MATS1v	2D autocorrelations	Moran autocorrelation − lag 1 / weighted by atomic van der Waals volumes
MATS1e	2D autocorrelations	Moran autocorrelation − lag 1 / weighted by atomic Sanderson electronegativities
MATS3e	2D autocorrelations	Moran autocorrelation − lag 3 / weighted by atomic Sanderson electronegativities
MATS5e	2D autocorrelations	Moran autocorrelation − lag 5 / weighted by atomic Sanderson electronegativities
MATS1p	2D autocorrelations	Moran autocorrelation − lag 1 / weighted by atomic polarizabilities
GATS4e	2D autocorrelations	Geary autocorrelation − lag 4 / weighted by atomic Sanderson electronegativities
EEig06x	Edge adjacency indices	Eigenvalue 06 from edge adj. matrix weighted by edge degrees
EEig07x	Edge adjacency indices	Eigenvalue 07 from edge adj. matrix weighted by edge degrees
EEig01r	Edge adjacency indices	Eigenvalue 01 from edge adj. matrix weighted by resonance integrals
ESpm01d	Edge adjacency indices	Spectral moment 01 from edge adj. matrix weighted by dipole moments
HOMA	Geometrical descriptors	Harmonic Oscillator Model of Aromaticity index
HOMT	Geometrical descriptors	HOMA total (trial)
G(N..N)	Geometrical descriptors	sum of geometrical distances between N..N
RDF105u	RDF descriptors	Radial Distribution Function − 10.5 / unweighted

RDF035m	RDF descriptors	Radial Distribution Function − 3.5 / weighted by atomic masses
RDF105v	RDF descriptors	Radial Distribution Function − 11.5 / weighted by atomic van der Waals volumes
RDF105e	RDF descriptors	Radial Distribution Function − 10.5 / weighted by atomic Sanderson electronegativities
RDF105p	RDF descriptors	Radial Distribution Function − 10.5 / weighted by atomic polarizabilities
Mor16u	3D-MoRSE descriptors	3D-MoRSE − signal 16 / unweighted
Mor16e	3D-MoRSE descriptors	3D-MoRSE − signal 16 / weighted by atomic Sanderson electronegativities
Ds	WHIM descriptors	D total accessibility index / weighted by atomic electrotopological states
HATS4m	GETAWAY descriptors	leverage-weighted autocorrelation of lag 4 / weighted by atomic masses
nRCONHR	Functional group counts	number of secondary amides (aliphatic)
nHDon	Functional group counts	number of donor atoms for H-bonds (N and O)
H–050	Atom-centered fragments	H attached to heteroatom
Hy	Molecular properties	hydrophilic factor
TPSA(NO)	Molecular properties	topological polar surface area using N, O polar contributions
TPSA(Tot)	Molecular properties	topological polar surface area using N, O, S, P polar contributions
B01[N–O]	2D binary fingerprints	presence/absence of N–O at topological distance 1
B06[O–O]	2D binary fingerprints	presence/absence of O–O at topological distance 6
F06[O–O]	2D frequency fingerprints	frequency of O–O at topological distance 6

Appendix: J. Features selected for Group II using PSA-SVM.

Name	Block	Description
Ss	Constitutional descriptors	sum of Kier–Hall electrotopological states
Ms	Constitutional descriptors	mean electrotopological state
nR05	Constitutional descriptors	number of 5-membered rings
CSI	Topological descriptors	eccentric connectivity index
ATS2p	2D autocorrelation indices	Broto–Moreau autocorrelation of a topological structure − lag 2 / weighted by atomic polarizabilities
MATS1p	2D autocorrelation indices	Moran autocorrelation − lag 1 / weighted by atomic polarizabilities
MATS3p	2D autocorrelation indices	Moran autocorrelation − lag 3 / weighted by atomic polarizabilities
GATS1v	2D autocorrelation indices	Geary autocorrelation − lag 1 / weighted by atomic van der Waals volumes
ESpm15u	Edge adjacency indices	Spectral moment 15 from edge adjacent matrix
BEHv5	Burden eigenvalue descriptors	highest eigenvalue n. 5 of Burden matrix / weighted by atomic van der Waals volumes
VRA1	Eigenvalue-based indices	Randic-type eigenvector-based index from adjacency matrix
G(O..Cl)	Geometrical descriptors	sum of geometrical distances between O..Cl
Mor27e	3D-MoRSE descriptors	3D-MoRSE − signal 27 / weighted by atomic Sanderson electronegativities
Tu	WHIM descriptors	T total size index / unweighted
R6p	GETAWAY descriptors	R autocorrelation of lag 6 / weighted by atomic polarizabilities
nArOH	Functional group counts	number of aromatic hydroxyls
nHBonds	Functional group counts	number of intramolecular H-bonds
F02[C–O]	2D frequency fingerprints	frequency of C–O at topological distance 2
F02[N–N]	2D frequency fingerprints	frequency of N–N at topological distance 2
F06[C–N]	2D frequency fingerprints	frequency of C–N at topological distance 6

Appendix: K. Features selected for Group III using PSA-SVM.

Name	Block	Description
CIC1	Information indices	complementary information content (neighborhood symmetry of 1-order)
ATS6m	2D autocorrelation indices	Broto–Moreau autocorrelation of a topological structure − lag 6 / weighted by atomic masses
ATS7p	2D autocorrelation indices	Broto–Moreau autocorrelation of a topological structure − lag 7 / weighted by atomic polarizabilities
MATS8e	2D autocorrelation indices	Moran autocorrelation − lag 8 / weighted by atomic Sanderson electronegativities
GATS8m	2D autocorrelation indices	Geary autocorrelation − lag 8 / weighted by atomic masses
ESpm14x	Edge adjacency indices	Spectral moment 14 from edge adjacent matrix weighted by edge degrees
QYYp	Geometrical descriptors	Qyy COMMA2 value / weighted by atomic polarizabilities
RDF020u	RDF descriptors	Radial Distribution Function − 2.0 / unweighted
Mor10e	3D-MoRSE descriptors	3D-MoRSE − signal 10 / weighted by atomic Sanderson electronegativities
G3u	WHIM descriptors	3rd component symmetry directional WHIM index / unweighted
HATS4u	GETAWAY descriptors	leverage-weighted autocorrelation of lag 4 / unweighted
R4u	GETAWAY descriptors	R autocorrelation of lag 4 / unweighted
C–037	Atom-centered fragments	Ar–CH=X (Ar represents aromatic groups; X represents any electronegative atom)
H–046	Atom-centered fragments	H attached to C (sp3) no X attached to next C
H–051	Atom-centered fragments	H attached to alpha–C (a C attached through a single bond with –C=X, –C#X, –C–X)
Hypertens-80	Molecular properties	Ghose–Viswanadhan–Wendoloski antihypertensive-like index at 80%
B10[C–N]	2D binary fingerprints	presence/absence of C–N at topological distance 10
F04[C–C]	2D frequency fingerprints	frequency of C–C at topological distance 4

Appendix: L. Features selected for Group IV using PSA-SVM.

Name	Block	Description
JhetZ	Topological descriptors	Balaban-type index from Z weighted distance matrix (Barysz matrix)
BAC	Topological descriptors	Balaban centric index
piPC10	Walk and path counts	molecular multiple path count of order 10
EEig01d	Edge adjacency indices	Eigenvalue 01 from edge adjacent matrix weighted by dipole moments
ESpm04u	Edge adjacency indices	Spectral moment 04 from edge adjacent matrix
BELm6	Burden eigenvalue descriptors	lowest eigenvalue n. 6 of Burden matrix / weighted by atomic masses
QXXe	Geometrical descriptors	Qxx COMMA2 value / weighted by atomic Sanderson electronegativities
RDF025e	RDF descriptors	Radial Distribution Function − 2.5 / weighted by atomic Sanderson electronegativities
Mor07u	3D-MoRSE descriptors	3D-MoRSE − signal 07 / unweighted
Mor12u	3D-MoRSE descriptors	3D-MoRSE − signal 12 / unweighted
Mor29m	3D-MoRSE descriptors	3D-MoRSE − signal 29 / weighted by atomic masses
Mor02v	3D-MoRSE descriptors	3D-MoRSE − signal 02 / weighted by atomic van der Waals volumes
Mor19p	3D-MoRSE descriptors	3D-MoRSE − signal 19 / weighted by atomic polarizabilities
Ks	WHIM descriptors	K global shape index / weighted by atomic electrotopological states
HATS0p	GETAWAY descriptors	leverage-weighted autocorrelation of lag 0 / weighted by atomic polarizabilities
nArOR	Functional group counts	number of ethers (Ar represents aromatic group; R represents any group)
C–024	Atom-centered fragments	R–CH–R (R represents any group linked through carbon)
Hypertens-80	Molecular properties	Ghose–Viswanadhan–Wendoloski antihypertensive-like index at 80%
F05[O–O]	2D frequency fingerprints	frequency of O–O at topological distance 5
F08[C–Cl]	2D frequency fingerprints	frequency of C–Cl at topological distance 8

Appendix: M. Features selected for Group V using PSA-SVM.

Name	Block	Description
BEHv1	Burden eigenvalue descriptors	highest eigenvalue n. 1 of Burden matrix / weighted by atomic van der Waals volumes
JGI8	Topological charge indices	mean topological charge index of order8
RDF085u	RDF descriptors	Radial Distribution Function − 8.5 / unweighted
RDF110v	RDF descriptors	Radial Distribution Function − 11.0 / weighted by atomic van der Waals volumes
RDF105p	RDF descriptors	Radial Distribution Function − 10.5 / weighted by atomic polarizabilities
Mor02m	3D-MoRSE descriptors	3D-MoRSE − signal 02 / weighted by atomic masses
Mor11v	3D-MoRSE descriptors	3D-MoRSE − signal 11 / weighted by atomic van der Waals volumes
Ts	WHIM descriptors	T total size index / weighted by atomic electrotopological states
F09[C–C]	2D frequency fingerprints	frequency of C–C at topological distance 9

Appendix: N. Features selected for Group VI using PSA-SVM.

Name	Block	Description
GMTIV	Topological descriptors	Gutman MTI by valence vertex degrees
ATS4v	2D autocorrelation indices	Broto–Moreau autocorrelation of a topological structure − lag 4 / weighted by atomic van der Waals volumes
EEig04d	Edge adjacency indices	Eigenvalue 04 from edge adjacent matrix weighted by dipole moments
EEig01r	Edge adjacency indices	Eigenvalue 01 from edge adjacent matrix weighted by resonance integrals
DP04	Randic molecular profiles	molecular profile no. 04
RDF010m	RDF descriptors	Radial Distribution Function − 1.0 / weighted by atomic masses
RDF070m	RDF descriptors	Radial Distribution Function − 7.0 / weighted by atomic masses
Mor26u	3D-MoRSE descriptors	3D-MoRSE − signal 26 / unweighted
Mor24m	3D-MoRSE descriptors	3D-MoRSE − signal 24 / weighted by atomic masses
P1e	WHIM descriptors	1st component shape directional WHIM index / weighted by atomic Sanderson electronegativities
nSO	Functional group counts	number of sulfoxides
H–051	Atom-centered fragments	H attached to alpha–C (a C attached through a single bond with –C=X, –C#X, –C–X)

Appendix: O. Features selected for Group VII using PSA-SVM.

Name	Block	Description
MSD	Topological descriptors	mean square distance index (Balaban)
BIC2	Information indices	bond information content (neighborhood symmetry of 2-order)
SIC3	Information indices	structural information content (neighborhood symmetry of 3-order)
RDF135v	RDF descriptors	Radial Distribution Function − 13.5 / weighted by atomic van der Waals volumes
RDF150v	RDF descriptors	Radial Distribution Function − 15.0 / weighted by atomic van der Waals volumes
Mor13u	3D-MoRSE descriptors	3D-MoRSE − signal 13 / unweighted
L1m	WHIM descriptors	1st component size directional WHIM index / weighted by atomic masses
nHDon	Functional group counts	number of donor atoms for H-bonds (N and O)
B10[N–S]	2D binary fingerprints	presence/absence of N–S at topological distance 10
B10[N–F]	2D binary fingerprints	presence/absence of N–F at topological distance 10
F06[N–N]	2D frequency fingerprints	frequency of N–N at topological distance 6

Appendix: P. Features selected for multi-class classification using RBE-RF.

Name	Block	Description
nCIR	Constitutional descriptors	number of circuits
RBN	Constitutional descriptors	number of rotatable bonds
nAB	Constitutional descriptors	number of aromatic bonds
TI2	Topological descriptors	second Mohar index TI2
Rww	Topological descriptors	reciprocal hyper-detour index
D/D	Topological descriptors	distance/detour index
D/Dr05	Topological descriptors	distance/detour ring index of order 5
D/Dr06	Topological descriptors	distance/detour ring index of order 6
SRW09	Walk and path counts	self-returning walk count of order 09
piPC05	Walk and path counts	molecular multiple path count of order 05
piPC06	Walk and path counts	molecular multiple path count of order 06
piPC08	Walk and path counts	molecular multiple path count of order 08
IDE	Information indices	mean information content on the distance equality
HVcpx	Information indices	graph vertex complexity index
EEig01x	Edge adjacency indices	Eigenvalue 01from edge adj. matrix weighted by edge degrees
EEig01r	Edge adjacency indices	Eigenvalue 01 from edge adj. matrix weighted by resonance integrals
BEHv1	Burden eigenvalue descriptors	highest eigenvalue n. 1 of Burden matrix / weighted by atomic van der Waals volumes
BEHv2	Burden eigenvalue descriptors	highest eigenvalue n. 2 of Burden matrix / weighted by atomic van der Waals volumes
BEHv3	Burden eigenvalue descriptors	highest eigenvalue n. 3 of Burden matrix / weighted by atomic van der Waals volumes
BEHe1	Burden eigenvalue descriptors	lowest eigenvalue n. 1 of Burden matrix / weighted by atomic Sanderson electronegativities
BEHe2	Burden eigenvalue descriptors	lowest eigenvalue n. 2 of Burden matrix / weighted by atomic Sanderson electronegativities
BEHe3	Burden eigenvalue descriptors	lowest eigenvalue n. 3 of Burden matrix / weighted by atomic Sanderson electronegativities
BEHp1	Burden eigenvalue descriptors	highest eigenvalue n. 1 of Burden matrix / weighted by atomic polarizabilities
BEHp2	Burden eigenvalue descriptors	highest eigenvalue n. 2 of Burden matrix / weighted by atomic polarizabilities
BEHp3	Burden eigenvalue descriptors	highest eigenvalue n. 3 of Burden matrix / weighted by atomic polarizabilities
LP1	Eigenvalue-based	Lovasz–Pelikan index (leading eigenvalue)

	indices	
E3u	WHIM descriptors	3rd component accessibility directional WHIM index / unweighted
nR=Ct	Functional group counts	number of aliphatic tertiary C(sp2)
nC=N–N<	Functional group counts	number of amidine derivatives
nPyrroles	Functional group counts	number of Pyrroles
nPyrimidines	Functional group counts	number of Pyrimidines
nHDon	Functional group counts	number of donor atoms for H-bonds (N and O)
C–017	Atom-centered fragments	=CR2 (R represents any group linked through carbon)
C–030	Atom-centered fragments	X–CH–X (X represents any electronegative atom (O, N, S, P, Se, halogens))
N–074	Atom-centered fragments	R#N/R=N– (R represents any group linked through carbon)
Hy	Molecular properties	hydrophilic factor
B01[N–N]	2D binary fingerprints	presence/absence of N–N at topological distance 1
F01[N–N]	2D frequency fingerprints	frequency of N–N at topological distance 1
F02[N–N]	2D frequency fingerprints	frequency of N–N at topological distance 2
F03[C–N]	2D frequency fingerprints	frequency of C–N at topological distance 3
F04[N–N]	2D frequency fingerprints	frequency of N–N at topological distance 4
F09[C–C]	2D frequency fingerprints	frequency of C–C at topological distance 5

VITA

Education

Ph.D. in Pharmaceutical Science, University of Mississippi

Dissertation: Discovery of Novel Glycogen Synthase Kinase-3β Inhibitors: Molecular Modeling, Virtual Screening, and Biological Evaluation.

M.S. in Computer & Information Science, University of Mississippi

Thesis: Implementation of Multiple-Instance Learning in Drug Activity Prediction: A Framework to Identify Bioactive Conformers.

M.S. in Medicinal Chemistry, Peking University Health Science Center, China

Thesis: Rational Design, Synthesis and Biological Evaluation of Novel Peptidomimetics for the Inhibition of Proteasome.

B.S. in Pharmaceutical Science, Peking University Health Science Center, China

Honors & Awards

University of Mississippi, Oxford, MS

- University of Mississippi Graduate Student Achievement Award (one of the two best graduate students in the School of Pharmacy at University of Mississippi in 2012)
- 2011 Nobles-Sam Graduate Research Award from the Department of Medicinal Chemistry at the University of Mississippi, bestowed on the student with the best podium presentation at the annual MALTO meeting (May 2011)
- 2011 Graduate Student Research Award presented by American Chemical Society the Ole Miss Local Section (April 2011)
- Research Symposium 2011 Best Poster Award in Pharmaceutical Sciences presented by the University of Mississippi Graduate Student Council (April 2011)
- Graduate Student Council Research Award presented by the Graduate School (Fall 2010)

- NIH Predoctoral Fellow 2009 sponsored by the Center of Research Excellence in Natural Products Neuroscience (CORE-NPN) (October 2009)

Peking University Health Science Center, Peking, China

- Peking University Yi Yao Scholarship; Awards for excellent student (2005-2006)
- Peking University GE Medical Education Scholarship; Awards for excellent student (2004-2005)
- Peking University Second-Class of Excellent Medicinal Student Scholarship (2003-2004)
- Dean’s Award for Physical Excellence (2001-2003)

List of Publications

1. Xiaofei Nan, Gang Fu, Zhendong Zhao, Sheng Liu, Ronak Y. Patel, Haining Liu, Pankaj R. Daga, Robert J. Doerksen, Xing Dang, Yixin Chen, Dawn Wilkins. “Leveraging Domain Information to Restructure Biological Prediction.” *BMC Bioinformatics*, 12, S22 (2011) 15, (2011 MCBIOS Proceedings).
2. Gang Fu, Robert J. Doerksen, and Ping Xu. “Assignment of absolute configuration of sulfinyl dilactones: Optical rotations and 1H NMR experiment and DFT calculations.” *Journal of Molecular Structure*, (2011) 987, 166-173.
3. Gang Fu, Xiao-Min Zou, Yi-Qiu Fu, De Mou, Chao Ma, Yang Lu and Ping Xu. “Synthesis of protected aminoalkyl sulfinyl dilactones from α-amino acids.” *Journal of Chinese Pharmaceutical Sciences,* (2007) 16, 119-124.
4. Man Xu, Gang Fu, Xue Qiao, Wan-Ying Wu, Hui Guo, Ai-Hua Liu, Jiang-Hao Sun, De-An Guo. “HPLC method for comparative study on tissue distribution in rat after oral administration of salvianolic acid B and phenolic acids from Salvia miltiorrhiza.” *Biomedical Chromatography*, (2007), 21, 1052-1063.
5. Man Xu, Zichuan Zhang, Gang Fu, Shifeng Sun, Jianghao Sun, Min Yang, Aihua Liu, Jian Han, Dean Guo. “Liquid chromatography–tandem mass spectrometry analysis of protocatechuic aldehyde and its phase I and II metabolites in rat.” *Journal of Chromatography B,* (2007), 856, 100-107.

6. Yiqiu Fu, Bo Xu, Xiaomin Zou, Chao Ma, Xiaoming Yang, Ke Mou, Gang Fu, Yang Lü, Ping Xu. “Design and synthesis of a novel class of furan-based molecules as potential 20S proteasome inhibitors.” *Bioorganic & Medicinal Chemistry Letters*, (2007), 17, 1102-1106.

List of Submitted Manuscripts

7. Gang Fu, Xiaofei, Nan, Haining Liu, Ronak Patel, Pankaj Daga, Yixin Chen, Dawn E. Wilkins, Robert J. Doerksen. “Implementation of multiple-instance learning in drug activity prediction to identify bioactive conformers” submitted to *BMC Bioinformatics* (2012 MCBIOS Proceedings).
8. Sheng, Liu, Ronak Y. Patel, Pankaj R. Daga, Haining Liu, Gang Fu, Robert J. Doerksen, Yixin Chen, and Dawn E. Wilkins. “Combined rule extraction and feature elimination in supervised classification” submitted to *IEEE Transactions on NanoBioscience* (2012).
9. Gang Fu, Prasanna Sivaprakasam, Olivia R. Dale, Susan P. Manly, Stephen J. Cutler, Robert J. Doerksen. “Pharmacophore modeling, ensemble docking and virtual screening studies on glycogen synthase kinase-3β.” submitted to *Journal of Chemical Information and Modeling* (2012).
10. Gang Fu, Haining Liu, and Robert J. Doerksen. “Induced fit docking, molecular dynamics simulations, binding energy calculations and QM/MM studies of the catalytic mechanism of human biliverdin IX$_\alpha$ reductase.” submitted to *Journal of Physical Chemistry* (2012).

List of Manuscripts in Preparation

11. Gang Fu, Robert J. Doerksen. “Recent advances in computational modeling and drug discovery aspects of glycogen synthase kinase-3” (Review article, in preparation).
12. Minkyun Ma, Joonseok Oh, In Hyun Hwang, Dong Woo Kim, Hiroyuki Osada, Jong Seog Ahn, Gang Fu, Robert J. Doerksen, Mark T. Hamann. “Diplostephiosides A and B, phenolic glycosides from the stems of the rare south american plant Diplostephium rhododendroides” To be submitted to *Biochimica et Biophysica Acta* (2012).
13. Gang Fu, Olivia R. Dale, Sheng Liu, Yixin Chen, Dawn E. Wilkins, Susan P. Manly, Stephen J. Cutler, Robert J. Doerksen. “Hierarchical quantitative structure-activity relationship and

virtual screening studies on glycogen synthase kinase-3*β*." To be submitted to *Journal of Chemical Information and Modeling* (2012).