Experimental and Computational Methods for Identification of Novel Fungal Histone Acetyltransferase Rtt109 Inhibitors

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Xia Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Elizabeth A. Amin, Advisor

February 2014

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Elizabeth A. Amin. She gave me the opportunities to explore and to learn in many areas of medicinal chemistry. I am indebted to her guidance, patience, understanding, and encouragement that have nurtured my academic growth at the University of Minnesota. I would also like to thank my committee members, Dr. Barry Finzel, Dr. Carston Wagner, and Dr. Douglas Ohlendorf for their insightful comments and suggestions that contribute heavily to this dissertation.

The research on Rtt109 presented in this dissertation would not have been possible without the help and support from the project team. I would like to specially thank Dr. Michael Walters for initiating the collaboration, answering my questions, and always providing useful suggestions. Thanks also to Jayme L. Dahlin, Rondedrick Sinville, Jonathan Solberg, and Derek Hook who conducted the high throughput screen and data analysis.

The Amin group has provided so much support over the last five years. The former member Dr. Richard Wood helped me get started. Dr. Ting-Lan Chiu always took the time to answer my questions and kept my spirits up when projects were difficult to continue. Thanks also to Elbek Kurbanov for interesting Discussion.

Finally, I would like to thank my family and friends. I am grateful that my parents, Rongrong Zhang and Yingjun Li, and my husband, Dawen Niu, have always supported and encouraged me. My special thanks go to my dear friends, Xing Liu, Xun Ming, Yanrong Zhu, Xiao Yi, Frank Chao, Yang Li, Ran Dai, Dan Wang, Wei Li, Yiwei

Zhang, Lipeng Ning, Yuqiang Qian, Ke Yang, Can Zhou, Wengui Zhang, Jing Yang, and Yuanyan Gu, Tao Wang, Xiangwei Tang, and Lei Dai. They have made my time in Minnesota enlightening and enjoyable.

Dedication

This thesis is dedicated to my husband, Dawen Niu, my son, Yi Niu, my father, Rongrong Zhang, and my mother, Yingjun Li.

Abstract

Rtt109 is a fungal-specific histone acetyltransferase that catalyzes histone H3 lysine 56 acetylation and is a promising antifungal drug target. To identify novel Rtt109 inhibitors as potential drug scaffolds, we employed *in vitro* high throughput screening (HTS) and various computer-assisted strategies, including molecular dynamics, docking, three-dimensional quantitative structure-activity relationship (3D-QSAR) analysis, pharmacophore modeling, and Support Vector Machine (SVM) mining. An initial experimental screening of 82,861 compounds (HTS1) yielded hits with activity ranging from 0.49 – 17.5 μM against Rtt109. The molecular dynamics simulation of Rtt109 suggested that the histone lysine tunnel, a potential inhibitor binding site, was not flexible and thus the use of a rigid protein structure of Rtt109 was appropriate for docking studies. From a virtual screen using Surflex-Dock, we have identified 878 additional compounds as potential hits, with predicted K_d values of 0.1 nm or lower. Based on preliminary experimental data from HTS1, validated pharmacophore maps were developed and used to pinpoint potential Rtt109 ligand-receptor interactions. 3D-QSAR CoMFA and CoMSIA models that were also derived from the hit series generated in the initial experimental HTS display high self-consistency ($r^2 = 0.985$ [CoMFA] and $r^2 = 0.976$ [CoMSIA]) and robust internal predictivity ($r_{cv}^2 = 0.754$ [CoMFA] and $r_{cv}^2 = 0.654$ [CoMSIA]). Importantly, key features identified in both the pharmacophore hypotheses and the 3D-QSAR models agreed well with each other and with experimentally defined structural features in the Rtt109 lysine-binding tunnel. In addition, our optimized SVM models demonstrated high predictive power against the external test sets for Rtt109 with

accuracy of 91.1%. We also identified novel features with significant differentiating ability to separate Rtt109 inhibitors from non-inhibitors.

Table of Contents

List of Tables

List of Figures

CHAPTER 1:

INTRODUCTION

1.1 Fungal Infections

Fungi are well known as a widespread threat to plant species.[1] Fungal infections in potatoes, rice, wheat, and soybean cause economic losses and present a growing threat to food security. Invasive tree diseases have led to the loss of ~100 million elm trees in the United States and the United Kingdom, with economic cost for the loss of wood and also increased carbon release.[2] In addition, fungal infections have caused population declines in bats, frogs and soft corals, bees, Tilapia fish, and so on. For example, during March 2007, bats in New York State died in great numbers because of the white nose syndrome caused by fungal infections.[3] The fungus responsible for causing the deadly white nose syndrome has also been found at Soudan Underground Mine State Park and Forestville/Mystery Cave State Park, two of Minnesota's largest bat caves.[4] Until 2007, some areas of central America had lost more than 40% of amphibian species due to fungal infections, leading to ecosystem-level changes.[5]

Fungal infections including systemic fungal infections (mycoses), also pose an increasing threat to human health, particularly in cases of immunocompromised individuals undergoing corticosteroid, immunosuppressant and/or antimetabolite therapy for organ transplantation, cancer, azotemia, diabetes mellitus, or AIDS.[6-8] Superficial infections which affect ~25% of the population worldwide are the most common human fungal diseases.[9] These infections often lead to athlete's foot, ringworm, and infection of

the nails. While superficial fungal infections are usually not life-threatening in immunocompetent hosts, systemic mycoses affecting severely compromised patients often exhibit acute presentations with rapidly progressive pneumonia and/or extrapulmonary dissemination. The mortality rates of patients can exceed 50% in the majority of systemic mycoses, and approach 100% in extrapulmonary invasive aspergillosis.[10-12] Although major advances have been made in combating fungal infections,[13] the design of fungal-specific therapeutics is not trivial because fungal pathogens are eukaryotic and share key biological processes with humans and other animals. Many currently available antifungal drugs exhibit an array of adverse effects due to lack of selectivity; in the case of amphotericin B, the standard therapy for most life-threatening systemic mycoses, these may include cardiac arrest, encephalopathy, neuropathy, renal damage, liver failure, thrombophlebitis and hearing loss.[14, 15] In addition, the rapid evolution of drug resistance has compromised the effectiveness of many currently available antifungal therapies, such as the widely used azole derivatives fluconazole and ketoconazole, that target lanosterol demethylase in the fungal ergosterol pathway.[14, 16-18] Undesirable drug-drug interactions due to interference with the CYP3A4, CYP2C9, and CYP2C19, also restrict the use of the azole antifungal agents.[19] Given the rising incidence of life-threatening systemic mycoses (mainly due to increasing numbers of immunocompromised individuals), there is a critical unmet need for new antifungal therapeutic treatments that selectively target fungi and can bypass current fungal resistance pathways.

1.2 Histone Acetyltransferase

Histones are the common components of chromosomes and are the most conservative proteins known. There are five types of histones: H1, H2A, H2B, H3, and H4. The histone H1 serves as a linker that interacts with DNA and stabilizes the highly ordered assembly of DNA helix and proteins.[20] One-hundred-and-forty-seven base pairs of DNA are wrapped around an octamer of core histones H2A, H2B, H3 and H4 to form the basic unit of chromatin, nucleosome.[21] Histones can undergo post-translational modifications to both their structured core domains and less structured tail domains, including more than 60 different residues on histones. At least eight different classes of modifications have been identified on histones: acetylation, methylation, phosphorylation, ubiquitylation, sumoylation, ADP ribosylation, deimination, and proline isomerization.[22] The dynamic modifications of histones are involved in the regulation of transcription, DNA repair, and chromosome condensation during different cell development stages.[23]

Acetylation is one of the most widely studied histone modifications. Histone acetyltransferases (HATs) are the enzymes responsible for the transfer of an acetyl group from AcCoA to histones. On the other hand, histone deacetylases (HDACs) catalyze the removal of acetyl groups so that histone acetylation is a reversible process.[24] Generally, acetylation occurs at the lysine residues of histones. The acetylation catalyzed by HATs results in the neutralization of the positively charged lysine residue of histone, which interrupts the interactions between the lysine residue and the negative charged DNA phosphate backbone.[25] As a result, chromatin is more extended and more accessible for

transcriptional factors, and gene transcription is promoted.

1.3 Histone Acetyltransferase Families

HATs are a diverse set of enzymes that can be classified into different families based on their catalytic domains and their sequence homology. The GNAT (Gcn5-related N-acetyltransferase) family is one of the main families, which includes Gcn5 (KAT2A), PCAF (KAT2B), Elp3 (KAT9), Hat1, Hpa2, and so on. HATs in this family are transcriptional activators and share more than 70% sequence homology. Another predominant family is the MYST family. The members of the MYST family are Morf (KAT6B), Ybf2 (Sas3), Sas2, Tip60 (KAT5), and so on. These HATs are involved in the DNA repair, gene anti-silencing, and dosage conpensation.[26] Other HATs include p300 (KAT3A), CBP (KAT3B), Taf1, and a number of steroid receptor co-activators such as p600 (KAT13C) and CLOCK (KAT13D). Like Gcn5, p300/CBP serve as co-activators for gene transcription. They interact with DNA-bound transcription factors including E1A and phosphorylated CREB, but not directly with DNA.[26, 27]

1.4 Rtt109

Rtt109 (Regulator of Ty1 Transposition 109) is a fungal-specific histone acetyltransferase that acetylates histone H3 lysine residues to promote gene activation and genome stability.[28] Gene deletion experiments[29, 30] have demonstrated that Rtt109 is essential for fungal pathogenicity: *C. albicans* Rtt109$^{-/-}$ mutant cells were found to be significantly less pathogenic in mice and more susceptible to *in vitro* phagocytosis than wild-type cells, as well as hypersensitive to genotoxic agents.[29] Rtt109 is structurally

similar to the mammalian HAT enzyme p300/CBP[31] but generally demonstrates little to no sequence homology with respect to other HATs. The specificity and catalytic activity of Rtt109 require a unique association with one of two histone chaperones, Vps75 and Asf1[16, 32, 33], an association not seen elsewhere in the HAT family. In the absence of chaperone proteins, Rtt109 is an inefficient histone acetyltransferase with a k_{cat} value of $2.3 \pm 0.7 \times 10^{-3}$ s^{-1} for histone H3.[34] When Rtt109 and Vps75 form a stable complex (K_d = 10 ± 2 nM), the k_{cat} value of the complex is 0.62±0.03 s^{-1} for full-length histone H3 and 0.11 ± 0.01 s^{-1} for H3 peptide substrate.[35] More detailed kinetic values for Rtt109 and Rtt109-Vps75 complex are summarized in Table 1.1, even though kinetic values vary slightly in assays done by different groups.[36] Further steady-state kinetic analyses indicate that Vps75 activates Rtt109 by enhancing catalytic turnover but not by increasing the binding affinity for histone. The enhancing rate of acetyl transfer occurs through the stabilization of the catalytically active conformation of Rtt109 by Vps75. Two electrostatic contact sites including residues Glu374, Glu378, Glu299, Glu300, and Asp301 of Rtt109 have been demonstrated to be important for Rtt109 catalytic activation. However, they are not the major reason for the high affinity between Rtt109 and Vps75. The hydrophobic interactions instead, likely involving the unstructured loop of Rtt109 (residues 130-179), mainly contribute to the overall high binding affinity.[34] For Rtt109-Asf1 complex, the k_{cat} value is 0.021 ± 0.002 s^{-1} and the k_{cat}/K_m value is $2.0 \pm 0.5 \times 10^4$ M^{-1} s^{-1} for histone H3/H4 substrates. Notably, unlike Rtt109-Vps75, Rtt109-Asf1 does not efficiently acetylate H3 but becomes efficient in the presence of H4, indicating the

interactions between Asf1 and H4 facilitate the proper binding or positioning of H3 for acetylation.[37]

Table 1.1. Kinetics of acetylation by Rtt109 and Rtt109-Vps75[36]

Enzyme	Substrate	k_{cat} (s^{-1})	K_m (μM)	k_{cat} / K_m (M^{-1} s^{-1})
Rtt109-Vps75	H3	0.21 ± 0.04	5.8 ± 0.8	$3.5 \pm 0.9 \times 10^4$
	H3-H4	0.11 ± 0.05	1.4 ± 0.4	$8.4 \pm 2 \times 10^4$
	H3 tail peptide	0.13 ± 0.04	75 ± 15	$1.8 \pm 0.5 \times 10^3$
	AcCoA	0.19 ± 0.01	1.0 ± 0.2	$1.9 \pm 0.4 \times 10^5$
Rtt109	H3	0.0033 ± 0.0003	8.1 ± 0.1	$4.3 \pm 0.2 \times 10^2$
	H3-H4	0.0044 ± 0.0009	2.9 ± 0.6	$1.5 \pm 0.6 \times 10^3$
	H3 tail peptide	0.0014 ± 0.0004	83 ± 29	$1.7 \pm 1 \times 10^1$
	AcCoA	0.0017 ± 0.0001	0.3 ± 0.09	$5.2 \pm 2 \times 10^3$

In addition, autoacetylation of Lys290 of Rtt109 is required for its catalysis. This autoacetylation stabilizes the interaction between the core protein acetyltransferase domain and the activation domain, and the engagement of the two domains turns Rtt109 into an active state.[38] Rtt109 catalyzes the autoacetylation of Lys290 itself by an intramolecular mechanism, meaning the autoacetylation occurs within a single enzyme molecule. Vps75 is not required for the autoacetylation and shows no apparent effects on the reaction rate and mechanism. Mutational studies demonstrate that autoacetylation stimulates histone acetyltransferase activity by enhancing catalytic acetyl transfer and increasing AcCoA binding. Both K290Q and K290R mutants, which represent unacetylated lysine, exhibit a ~100-fold decrease in k_{cat} of the Rtt109-Vps75 complex and a larger than 10-fold decrease in K_d value for AcCoA binding.[39]

The catalytic cores, catalytic mechanisms, and substrate specificity of Rtt109 also differ significantly from those of other known HATs including p300/CBP.[40] Unlike the Gcn5/PCAF HAT group, Rtt109 does not contain a key conserved active-site glutamate residue;[16] two important catalytic residues in p300/CBP (Trp1436 and Tyr1467) are also absent in Rtt109. As it is currently understood, the Rtt109 mechanism of action does not follow the ordered Bi-Bi ternary complex mechanism utilized by Gcn5/PCFA[33], the ternary ping-pong mechanisms proposed for Esa1 (a member of the MYST family of HATs)[28], or the Theorell-Chance mechanism for p300/CBP.[41] Rtt109 instead employs a random sequential mechanism in which the histone H3 and AcCoA moieties bind to the Rtt109-Vps75 complex in no particular order, followed by direct attack of the unprotonated histone lysine on AcCoA.[42] Finally, Rtt109 is quite selective for the histone

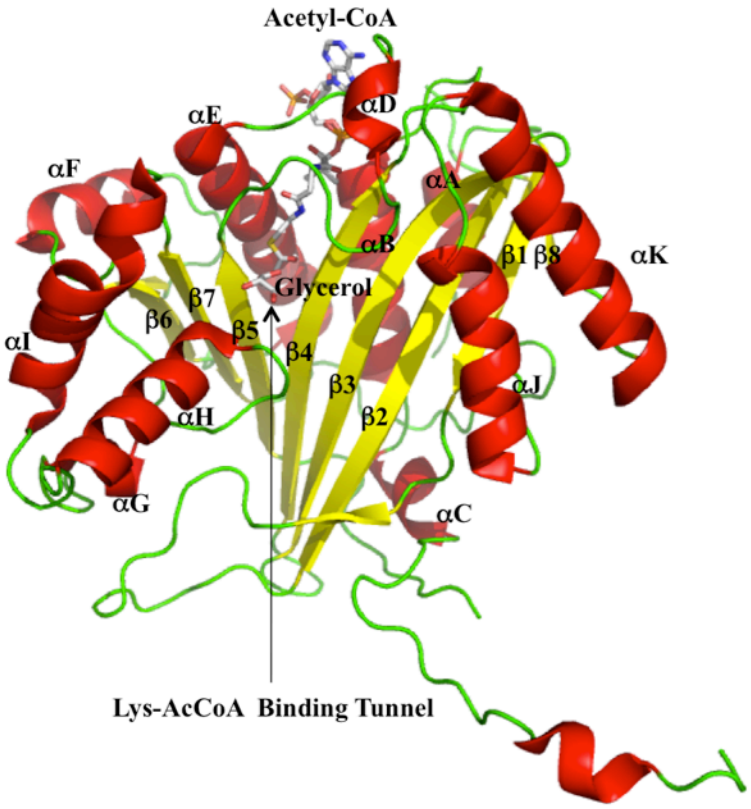
H3 lysines 56[32], 9[43], and 27[34], while other HATs including p300/CBP exhibit a much broader range of substrates (several lysine sites within histone and non-histone proteins).[42] Taken together, these key structural and mechanistic variances set Rtt109 apart from human HATs, and indicate that Rtt109 may be a promising target for small-molecule antifungal drug development. However, no Rtt109 inhibitors are currently available as therapeutics; in fact, recent *in vitro* studies have shown that a series of potent p300/CBP inhibitors exhibited only very weak activity against Rtt109.[31]

The structure of Rtt109 consists of eight β-strands surrounded by eleven α-helices and several loops (Figure 1.1). The Rtt109 surface features a roughly circular aperture located between the β-strands β4 and β5, and the surface loops β5-αE and αG-αH.[5, 16] This aperture leads into a hydrophobic tunnel (the histone lysine residue entrance tunnel), ending in the acetyl group of AcCoA when bound. On the opposite side of the enzyme, the AcCoA binding tunnel is positioned between the N-terminal areas of helices αB and αE, and the C-terminal regions of β-strands β4 and β5. In the histone lysine binding tunnel, Ser86, Tyr199, Trp222, and Asp288, which are proximal to the acetyl group of AcCoA, may function catalytically,[31] while Phe84, Tyr199, Pro289, and the sidechains of Lys87 and Arg194 may engage in hydrophobic interactions with the aliphatic part of the substrate histone lysine residue.[31] Single-site mutagenesis studies have shown that residues Asp89, Tyr199, Trp222, Asp287 and Asp288 are important for Rtt109 catalysis and/or substrate binding,[31, 38, 44] suggesting that these residues may be suitable as initial targets for inhibitor design. W222F and D89N reduce the apparent K_m for both H3 and

AcCoA, which leads to a ~25-fold rate decrease. Y199S results in a rate decrease of about 10-fold, and D287A has a mild 3-fold rate decrease.[47]

So far, not much work has been reported regarding inhibiton pf Rtt109, but there is considerable precedence for inhibition of other HATs. Thus, Rtt109 might be a promising target for antifugal agents.

Figure 1.1. Structure of fungal histone acetyltransferase Rtt109 (PDB code 2ZFN[16]) colored by secondary structure (PyMOL/1.4[45])



1.5 Histone Acetyltransferase Inhibitors

Dysfunction of HATs is often associated with the development of cancer, inflammation, deregulation of neuronal cell plasticity, and so on. In addition, the acetylation of HIV Tat protein is found to be necessary for viral pathogenesis, and acetylation of histone H3 lysine 56 in *Candida albicans* helps maintain fungal genome stability.[24] Thus, HATs are important targets for the development of novel therapeutics for cancer, Alzheimer disease, AIDS, and fungal infections. Investigations of HAT inhibitors as potential therapeutical strategies have been reported recently. Although various approaches including high-throughput screening and structure-based design have been utilized to discover novel inhibitors of HATs, a limited number of inhibitors have been identified. The available inhibitors include natural products, bisubstrate inhibitors based on AcCoA, and synthetic compounds. In the work described in this disseration, we are particularly interested in identifying and developing inhibitors against fungal specific histone acetyltransferase as potential antifugal agents, because it has been demonstrated that fungal histone acetyltransferase Rtt109 is essential for fungal pathogencity.

Glycosaminoglycans (GAGs) were reported to be potent inhibitors of p300 and PCAF *in vitro*. Heparin from the GAGs class may serve as an endogenous HAT inhibitor to modulate HAT activities. Heparin was shown to cause a significant increase in the apparent K_m of PCAF for H4 (5.3 ± 1.9 nM in the absence of heparin and 259.7 ± 79.8 nM in the presence of heparin) and further analysis suggests that heparin inhibits PCAF through a competitive-like mechanism with an apparent K_i of ~15 nM.[46] Another potential endogenous HAT inhibitor is spermidine, a natural polyamine. Administration

of spermidine leads to deacetylation of histone H3 via inhibition of histone acetyltransferases (HAT). In addition, depletion of endogenous polyamines results in hyperacetylation, increased oxidative stress, necrosis, and decreased lifespan.[47]

The natural product anacardic acid **1.1** (Figure 1.2) is an ingredient found in cashew nut shells that is associated with antitumor activity. It is reported to inhibit HAT activity of PCAF and p300. Its IC_{50} values against PCAF and p300 are ~5 μM and ~8.5 μM, respectively. An inhibition kinetics study suggests that anacardic acid is a non-competitive inhibitor of p300.[48] Anacardic acid also inhibits Tip60 *in vitro* with an IC_{50} of 9 μM. It blocks the Tip60-dependent activation of ATM protein kinase that is essential for cells to repair and to survive exposure of ionizing radiation and thus sensitizes tumor cells to the cytotoxic effects of ionizing radiation.[49] Although it is a non-specific inhibitor of HATs, anacardic acid provides a template for further modification and development of more specific inhibitors of HATs.

Curcumin **1.2** (Figure 1.2) is a polyphenolic compound from *Curcuma longa* rhizome. Two different HAT assays conducted by Balasubramanyam and co-workers demonstrated that the acetylation of histones H3 and H4 by p300/CBP was inhibited by curcumin with an IC_{50} of ~25 μM. However, it did not change the PCAF activity even at the concentration of 100 μM and thus shows a certain level of selectivity in the HAT family.[50] Enzyme kinetics studies indicate that curcumin does not bind to the active sites of histone or AcCoA. It instead binds to some allosteric site, which triggers a conformational change of p300, leading to a decrease of binding efficiency of histones and AcCoA. Curcumin also inhibits p300-mediated acetylation of p53 *in vivo*, but other

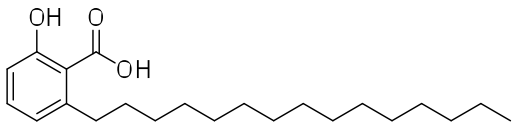
HAT-mediated acetylations of p53 are not inhibited by cucumin. In addition, curcumin inhibits the acetylation of the HIV-1 transactivator Tat that is important for the HIV transcriptional activation and suppresses the proliferation of HIV virus. Thus, curcumin may serve as a lead compound in the development of HIV therapeutics. So far, curcumin has been evaluated in several clinical trials for the treatment of myeloma, pancreatic cancer, chronic psoriasis vulgaris, Alzheimer's disease, and so on. It is the first small molecule with HAT inhibitory activity in clinical studies.[51]

Garcinol **1.3** (Figure 1.2), a polyisoprenylated benzophenone derivative from the fruit of *Garcinia indica*, has been shown by Balasubramanyam and co-workers to inhibit PCAF with an IC_{50} of ~5 μM and p300 with an IC_{50} of ~7 μM. It inhibits the HAT activity-dependent chromatin transcription, but not the transcription from a DNA template. In addition, garcinol stimulates apoptosis in HeLa cells and human leukemia cell lines. The kinetic analysis demonstrates that the inhibition mechanisms of garcinol for p300 and PCAF are similar. It behaves as an uncompetitive inhibitor for AcCoA, and as a competitive inhibitor for core histones.[52] However, interestingly, in another kinetic study done by Arifand co-workers garcinol showed competitive inhibition for the AcCoA binding site on p300 and noncompetitive inhibition for the histone-binding site.[53] A later isothermal calorimetric experiment then suggested garcinol binds to two sites on p300. Based on further studies using fluorescence, docking, and mutations, the authors proposed that the catechol hydroxy moieties provide key interactions with the AcCoA binding site and the isoprenoid groups play an important role in the binding to an allosteric site.[53]

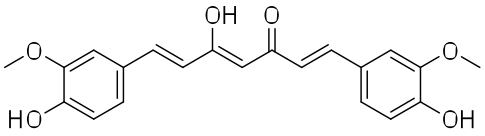
Epigallocatechin-3-gallate (EGCG) **1.4** (Figure 1.2), the major polyphenol found in green tea extract, has been reported to be an inhibitor of p300 and CBP with IC_{50} values of ~30 μM and ~50 μM, respectively. It also shows inhibitory activities against PCAF (IC_{50} of ~60 μM) and Tip60 (IC_{50} of ~70 μM).[54] Kinetic studies demonstrate that EGCG does not bind to the active site of histone and it uncompetitively inhibits p300/CBP. EGCG prevents the hyperacetylation of p65 in vitro, and represses tumor necrosis factor α (TNFα)–induced NF-κB activation. EGCG also reduces the release of IL-6 by cell cultures exposed to different inflammatory stimuli. It generally inhibits the inflammatory responses and thus may contribute to the prevention and treatment of diseases such as chronic obstructive pulmonary disease and asthma.[54]

Plumbagin (PTK-1) **1.5** (Figure 1.2), isolated from *Plumbago rosea* root extract, potently inhibits p300 *in vivo*, but not PCAF. It inhibits the HAT activity of the catalytically active domain with an IC_{50} of ~2 μM while inhibiting the full-length p300 with an IC_{50} of ~20 μM. The kinetic analysis of the inhibition of p300 by PTK-1 shows a noncompetitive pattern with both histones and AcCoA. Docking and mutational studies conducted by Ravindra and co-workers demonstrated that the hydroxyl group of PTK-1 was essential for its HAT inhibition and it was proposed that it forms a H-bond to the key residue K1358 in p300.[55] Further structure-activity relationship studies illustrated that the absence of a hydroxyl group led to inactive derivatives of PTK-1.

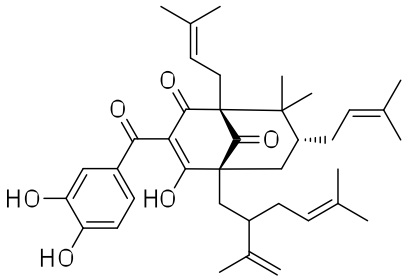
Figure 1.2. Natural products with HAT inhibitory properties



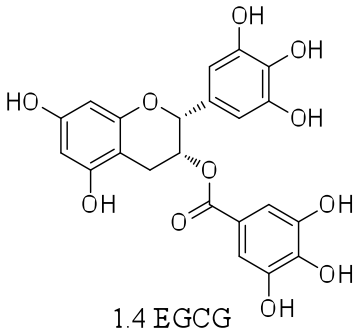
1.1 anacardic acid



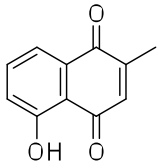
1.2 curcumin



1.3 garcinol



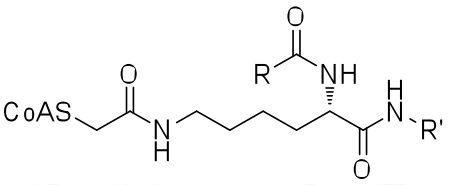
1.4 EGCG



1.5 plumbagin

Bisubstrate inhibitors are the first series of synthetic HAT inhibitors. These inhibitors are often derivatives of CoA, covalently connected to the substrate peptides with different chain lengths. An acetyl bridge between the amine substrate and CoA has been demonstrated by Lau and co-workers to be among the most effective linkers between the two and thus most of these bisubstrate inhibitors are designed with the acetyl bridge. Lys-CoA **1.6** (Figure 1.3) is one of the bisubstrate inhibiors that inhibit p300 selectively with an IC_{50} of ~0.5 μM, whereas its IC_{50} for PCAF is ~200 μM.[56] H3-CoA-20 **1.7** (Figure 1.3) proved to be potent at inhibiting PCAF with an IC_{50} of ~0.3 μM, but showed little potency against p300 with an IC_{50} of ~200 μM. Different from expected, neither CoA-SH or H3-20 peptide is a potent HAT inhibitor. H4K16CoA **1.8** (Figure 1.3) was one of the most potent inhibitor of Tip60 with an IC_{50} of ~17.6 μM. It also shows inhibitory activities against Esa1 with an IC_{50} of ~5.5 μM, against p300 with an IC_{50} of ~6.6 μM, and against PCAF with an IC_{50} of ~58.5 μM. The kinetic studies indicate that H4K16CoA is a competitive inhibitor of Esa1 versus AcCoA and noncompetitive inhibitor versus H4 peptide substrate.[57] Spd-CoA **1.9** (Figure 1.3) is another p300/CBP inhibitor. It inhibits histone acetylation, DNA synthesis, and acetylation-dependent DNA repair. In order to increase cell uptake, compound **1.10** was designed with truncated CoA moiety to eliminatethe negative charges. It is slightly more potent against p300/CBP *in vitro* than compound **1.9,** and it also suppresses histone acetylation *in vivo*.[58]

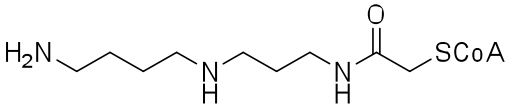
Figure 1.3. Bisubstrate HAT inhibitors



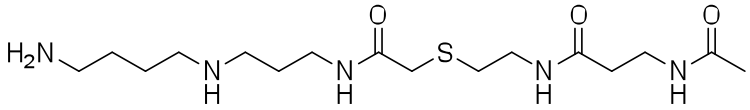
1.6 Lys-CoA $R = CH_3$, $R' = H$

1.7 H3-CoA-20 $R = G\text{-}G\text{-}T\text{-}S\text{-}L\text{-}R\text{-}A\text{-}T\text{-}Q\text{-}K\text{-}T\text{-}R\text{-}A\text{-}NHCH_3$, $R' = A\text{-}P\text{-}R\text{-}K\text{-}Q\text{-}L\text{-}OH$

1.8 H4K16CoA $R = R\text{-}H\text{-}R\text{-}K\text{-}NH_2$, $R' = K\text{-}A\text{-}G\text{-}G\text{-}K\text{-}G\text{-}L\text{-}G\text{-}K\text{-}G\text{-}G\text{-}K\text{-}G\text{-}R\text{-}G\text{-}S\text{-}OCH_3$



1.9 Spd-CoA



1.10

The γ-butyrolactone scaffold was chosen by Biel and co-workers for the development of new HAT inhibitors because of its recurrent structure in many natural products with a wide spectrum of biological activities. The α-methylene-γ-butyrolactone derivatives **1.11 -1.13** (Figure 1.4) display weak inhibition of CBP (IC_{50}values ranging from 0.5mM to 2mM), whereas compound **1.11** shows stronger inhibitory activity against Gcn5 with an IC_{50} value of 100 μM. Based on the kinetic studies, the inhibition of Gcn5 by compound **1.11** shows no time dependence. Thus compound **1.11** is a reversible inhibitor of Gcn5. A Michael addition of nucleophilic groups of the Gcn5 binding site to the α, β-unsaturated carbonyl moiety of compound **1.11** is unlikely to occur.[59]

The quinoline derivatives **1.14** and **1.15** (Figure 1.4) were designed by Mai and co-workers using anacardic acid as the template. Both compounds are able to reduce the yeast cell growth and the Gcn5-dependent gene transcription both in basal and in activated conditions. The histone H3 acetylation levels are highly decreased by the treatment of 0.6 mM **1.14** and 1.5 mM **1.15**.[60]

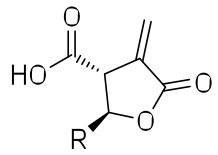
Thirty-five N-substituted isothiazolones were identified by Stimson and co-workers as p300 and PCAF inhibitors via a FlashPlate high throughput screening.[61] Compound **1.16** (CCT077791) and Compound **1.17** (CCT077792) reduced cellular acetylation in a time-and-concentration dependent manner in human colon tumor cell lines. The chemical reactivity of these compounds is related to their HAT inhibitory activity, as indicated by the loss of such activity in the presence of DTT. It is proposed that isothiazolones react with the side chain of cysteine groups to form a new covalent bond with consequent loss of catalytic activity of the HATs. The high reactivity and

considerable off-target effects of these compounds may restrict their use as therapeutics. However, they provide a starting point for the development of related inhibitors.

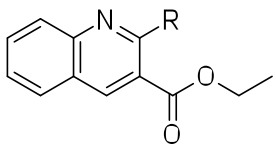
Structure-based *in silico* screening conducted by Bowers and coworkers led to the discovery of the compounds C464 **1.18**, C375 **1.19**, and C146 **1.20** (Figure 1.4) as p300 inhibitors.[62] About 500,000 commercially available compounds were docked into the lysine-CoA binding site of the p300 crystal structure and 194 highest-scoring compounds were purchased and tested experimentally. Among the 194 compounds, compound **1.18**, **1.19**, and **1.20** display relatively potent inhibitory activity against p300. The specificity of each of the three compounds were then analyzed versus six other HATs including Gcn5, PCAF, MOZ, and Rtt109. Compound **1.18** was highly selective in inhibiting p300 while compound **1.19** and compound **1.20** were less selective. Both compounds **1.19** and **1.20** showed inhibitory activity against at least one of the other HATs with comparable potency to that of p300. In the kinetic studies, compound **1.18** showed competitive inhibition of p300 versus AcCoA with a K_i value of 400 nM and a noncompetitive pattern versus H4-15 peptide substrate. Compound **1.19** proved to be a classical noncompetitive inhibitor of p300 with a K_i of 4.8 μM. Compound **1.20** was found to be a competitive inhibitor of p300 versus H4-15 peptide substrate. On the other hand, compound **1.20** was a pseudocompetitve inhibitor versus AcCoA with a K_i of 4.7 μM, indicating a more complex interaction with p300 than a bisubstrate analog. In the computational model, compound **1.18** was predicted to H-bond with the side chains of Thr1411, Tyr1467, Trp1466, and Arg1410 in the lysine-CoA binding site. Site-directed p300 mutations of each of these residues resulted in at least 2-fold increase in the apparent K_i of compound

1.18 and confirmed the binding mode of compound **1.18**. Further structure-activity relationship investigations demonstrated that the shape and/or electronic properties of the conjugated system in compound **1.18** were important for its activity. Replacing the para-carboxylic acid group of compound **1.18** with other hydrogen bond acceptors was well tolerated at both the *para*- and *meta*- positions. The nitrophenyl group could be replaced by a methylbenzoate or a cyanophenyl functionality. However, replacement of nitrophenyl moiety with a pyridine ring led to loss of its inhibition against p300.

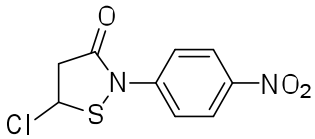
Figure 1.4. Synthetic small molecules as HAT inhibitors



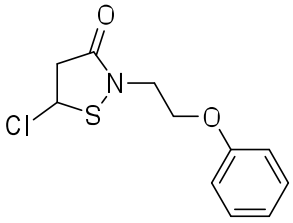
1.11 R = *n*-C_3H_7
1.12 R = *n*-C_4H_9
1.13 R = *n*-C_5H_{11}



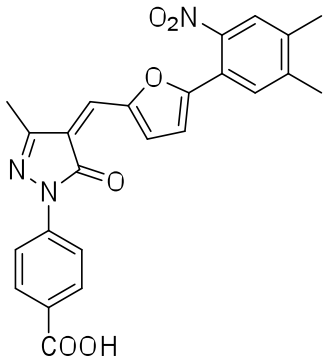
1.14 R = H
1.15 R = CH_3



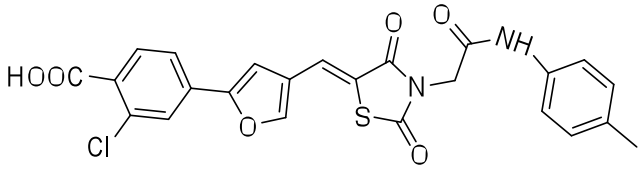
1.16 CCT077791



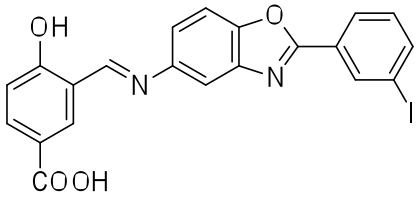
1.17 CCT077792



1.18 C646



1.19 C376

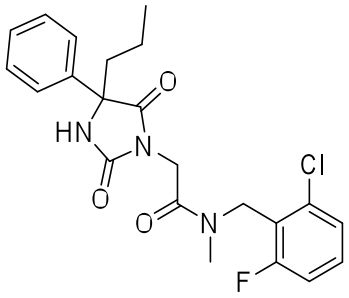


1.20 C146

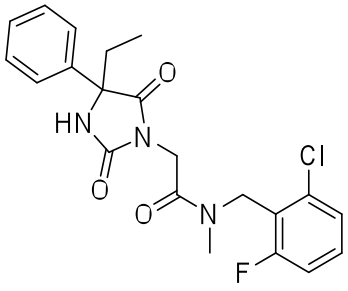
1.6 Development of Rtt109 Inhibitors

Besides our own efforts to be described, the Kaufman group has also contributed to the discovery and development of novel Rtt109 inhibitors. Kaufman and co-workers identified 537 initial hits with percentage inhibition larger than 50% at 25 μM from 363,843 molecules using a fluorescence-based assay.[63] Among them, 449 molecules were re-tested in an 8-point, 2-fold dose titration using optical absorbance detection, which was designed to eliminate non-specific fluorescence quenchers. Eighty-two compounds had a dose response value with 50% inhibition less than or equal to 20 μM; 28 of them were re-tested from powder samples. After the re-testing, nine compounds with a 50% inhibition value of less than 10 μM were further tested for their specificity. Among them, only compound **1.21** (Figure 1.5) inhibited Rtt109 without significantly inhibiting mammalian p300 or Gcn5. Its IC_{50} value is about 56 nM and it inhibited acetylation by both Rtt109-Asf1 and Rtt109-Vps75 complexes in similar ways. Kinetic studies showed that compound **1.21** reduced V_{max} and K_m for both AcCoA and the H3 histone peptide. Further studies suggested that compound **1.21** is not a competitive inhibitor of either AcCoA or the H3 peptide, but a tight or perhaps irreversible inhibitor that displays time-dependent inhibition against Rtt109. The following structure-activity relationship studies illustrated a very narrow SAR for compounds within the same chemical class as compound **1.21**. Small structural changes of compound **1.21** led to loss of activity or low activity. Among the structurally related compounds, only one compound **1.22** displayed a poor dose-response and minimal inhibition at the highest concentration tested (1 μM).

Figure 1.5. Small molecules as Rtt109 inhibitors



1.21



1.22

1.7 Concluding Remarks

Infectious diseases caused by fungi are presenting a worldwide threat not only to animals and plants but also to human health. However, because of the adverse effects, drug-drug interactions, and drug resistance, the use of many currently available antifungal therapies is restricted. New antifungal treatments that can overcome the listed deficiencies are thus desperately needed.

HATs are enzymes that acetylate histones. They are often involved in the regulation of transcription, DNA repair, and other gene regulations. Many HATs have been demonstrated to be important drug targets. Among them, Rtt109, a fungal-specific histone acetyltransferase, has been shown as a potential antifungal therapeutic target. Rtt109 has several unique features. Its catalytic cores, catalytic mechanisms, and substrate specificity differ significantly from other HATs. Additionally, the specificity and catalytic activity of Rtt109 require an association with one of two histone chaperones, Vps75 and Asf1. The availability of Rtt109 X-ray crystal structure enables structure-based inhibitor design. However, only a limited number of inhibitors were identified even for the whole HAT family. Most of these HAT inhibitors are natural products or natural product derivatives including garcinol, curcumin, and Lys-CoA that lack selectivity and demonstrate only moderate potency. Thus, there is a critical unmet need for potent and selective Rtt109 inhibitors. Further studies on development of novel Rtt109 inhibitors as pharmacological tools will shed the light on the understanding of roles of Rtt109 in fungal pathogenesis and will also ultimately lead to effective antifungal therapeutics for diseases such as candidemia and aspergillosis.

CHAPTER 2:

EXPERIMENTAL HIGH-THROUGHPUT SCREENING FOR INHIBITORS OF RTT109

2.1 Introduction

High-throughput screening (HTS) is a widely used technique in the drug discovery process. By using automatic systems, HTS allows researchers to quickly perform bioassays to provide probe or lead compounds that facilitate scientific research and drug discovery.[64] Many successful drug-discovery examples that utilize HTS can be found in the literature, including the discovery of the anti-HIV drug Maraviroc, a chemokine receptor antagonist, eltrombopag, a thrombopoietin receptor agonist, and the hepatitis C virus NS5A inhibitors.[65] Compared to other approaches such as structure-based and ligand-based drug design, one advantage of HTS is that the specific structural information of the target, mechanism of action, or information on other active compounds are not necessary for identifying compounds that modulate biological activity. However, there are also several limitations or drawbacks associated with HTS. Usually every compound is only tested at a single concentration in the primary screening, which may lead to false negatives. Depending on the components used in the bioassay and the detection methodology, specific classes of false positives may appear as true actives in the initial screening. These false positives need to be identified with counter screens to reduce the occurrence of misleading data. Another drawback of HTS is that it costs a significant amount of money, which may cause a financial burden in academic settings.

In our studies, *in vitro* HTS was used to identify Rtt109 inhibitors. The issues

associated with HTS mentioned above will also be discussed in the following sections.

2.2 Methods and Results

In vitro Assays. High-throughput screening kinetic assays were performed following Trievel et al[66] and Chung et al.[67] 15 nL DMSO or 10 μM test compounds were added to wells in Corning 384-well plates (Cat. No. 3677) using an ECHO 550 contactless liquid handler (Labcyte, Sunnyvale, CA). Controls were added to the appropriate wells using the same dispensing method. Next, 5 μL 1X HAT buffer (50mM Tris-HCl, 0.1mM EDTA, 50mM KCl, pH 8.0) was added using a Thermo Scientific (Rockford, IL) Multidrop 384 bulk reagent liquid dispenser. Following this, the Rtt109-Vsp75 complex (200 ng/well) and then Asf1-dH3/H4 (800 ng/well) were added using this Multidrop dispenser. To initiate the reaction, 10 μL of 15 μM acetyl-coenzyme A was added to the entire plate and incubated at 30°C for 1-2 hours. Following incubation, 5 μL 80 μM 7-diethylamino-3-(4'-maleimidylphenyl)-4-methylcoumarin (CPM, final concentration 20μM) that react with the free coenzyme A (CoASH) was added. The fluorescence intensity was measured using an M2E plate reader (Molecular Devices) with excitation at 405 nm and emission at 530 nm. The assay was validated for top-to-bottom and edge-to-edge variability using a control plate where column1 included all components (Rtt109-Vsp75, Afs1-H3/H4, AcCoA), column 2 lacked Rtt109-Vsp75, and column 3 lacked Afs1-H3/H4. The z-prime values ranged from 0.51-0.67. Finally, the HTS protocol was validated using duplicated assay runs of the LOPAC compound collection (Sigma-Aldrich, St. Louis, MO) on two separate days. The LOPAC results

indicated good reproducibility from day to day and from plate to plate and identified the same active compounds in each of the four replicate plates assayed with similar or better z-prime values than the control validation plates. Further details regarding assay development are being published under separate cover.[68]

Two Production Screenings. Due to the different purposes and priorities in different research phases, 82,861 compounds and 142,841 compounds from the University of Minnesota in-house "Gopher" (GPHR) compound library were screened separately in two production runs. The first HTS (HTS1) was conducted with a detergent-free buffer. Its aim was to quickly identify potential Rtt109 inhibitors. However, after HTS1 was finished, a preliminary analysis of the structural properties of the compounds showing apparent inhibition against Rtt109 suggested that potential chemical aggregation may lead to the inhibitor-like signals during HTS experiments.[69] Thus, based on the experience gained in the HTS1, 0.01% Triton X-100 was included in the second HTS (HTS2) to reduce the false positives due to non-specific chemical aggregation.[70] 667 compounds and 919 compounds were identified as actives from HTS1 and HTS2 respectively, as they exhibited a percent inhibition greater than three standard deviations (3σ) above the mean. These active compounds were cherry-picked for re-testing and were evaluated for activity using the same assay and an 8-point, 5-fold dilution dose-response curve. Known HATs inhibitors garcinol, curcumin, and a curcumin analog 2,6-*bis*-(3-bromo-4-hydroxybenzylidene)cyclohexanone served as positive controls while fluconazole with no reported activity against HATs was used as a negative control. All of the positive controls displayed dose-dependent inhibitory signals in the assays. An

orthogonal antibody-based slot blot assay was then applied to confirm that the inhibitory signals corresponded to the true decrease in histone acetylation catalyzed by Rtt109. The slot blot assay was performed following standard procedures using a Bio-Rad Bio-Dot SF microfiltration apparatus. The resulting membranes were imaged using a LI-COR Odyssey andanalyzed using Image Studio (LI-COR Biosciences). Additional control experiments were included where only the Rtt109-Vps75 and Asf1-dH3-H4 complexes and no AcCoA were added to the plate. Slot blot assays for each compound were replicated at least three times by independent experiments. All three positive controls exhibited dose-dependent decreases in H3 lysine 56 acetylation in the slot blot assays, while fluconazole did not show any detectable decrease in the histone acetylation.

Interference Assays. To eliminate false positives that produce inhibitory signals by reacting with the reagents in the *in vitro* HTS assay, interference assays were developed. The assay conditions and procedures were similar to that in the *in vitro* HTS assay except that no proteins were added to the reaction mixture. The first interference assay was designed to test the signal attenuation without proteins. Compounds that do not interfere with reagents CPM and CoASH should have similar signals compared to the DMSO control. Compounds that interfere with reagents will likely have diminished signal relative to the control. The second interference assay was used to measure the fluorescence of the compound itself and detect the formation of fluorescent adducts. Compounds that fluoresce themselves under the HTS conditions or produce fluorescent adducts with CPM, AcCoA, or CoASH may interfere with the HTS assay and may require additional assays to confirm their activities.

2.3 Discussion

The careful design and analysis by our HTS team led to a HTS assay for the Rtt109 complex that was more physiologically relevant compared to other reported Rtt109 HTS assays. Full-length histone proteins instead of histone peptides were incorporated in this assay. Both chaperones Asf1 and Vps75 were also included because of their contributions to the catalytic rate and stability of Rtt109. Thus, our HTS assay should reduce the probability of generating misleading data that are due to the use of highly artificial and non-physiological assay and would allow the identification of a broader spectrum of inhibitors with varying mechanisms of action compared to the reported Rtt109 HTS assays. For example, inhibitors identified by our HTS assay can directly target Rtt109 or disrupt protein-protein interactions between Rtt109 and Vps75, between Rtt109-Vps75 and its substrate histone, or interactions between Asf1 and histone H3-H4. Inhibitors may also work by binding to a selection of allosteric sites and thereby changing protein conformations affecting histone acetylation. The exact Rtt109 inhibitory mechanism has not been definitively established and additional mechanistic experiments would be helpful. In addition, HTS assays in general may return false-positives due to chemical aggregation, redox activity, and non-specific reactivity. Follow-up experiments may be able to triage these false-positives.

The HTS1 experiments measured Rtt109 activity only and did not correct for false positives. However, a clear set of actives and inactives was identified which provided training sets for the modeling experiments described below. Although the subsequent data from HTS2 differ from those from HTS1, all models were validated internally and

models obtained using different techniques agreed well with each other and yielded results strongly congruent with the experimentally defined structural features in the Rtt109 lysine binding tunnel. These models will be useful to guide the design of small molecules targeting the lysine tunnel.

CHAPTER 3:

MOLECULAR DYNAMICS SIMULATION ON RTT109

3.1 Introduction

Proteins are in constant motion in physiological environments.[71] Experimental studies have demonstrated that protein flexibility plays an important role in protein's biological effects, ligand binding, binding orientation, binding kinetics, and so on. Thus, understanding the molecular motions of proteins is undoubtedly crucial to drug design. However, experiments such as crystallographic and NMR studies that help illustrate protein motions are expensive and require extensive labor. Researchers have been seeking the relatively cheap computational tools to predict protein dynamics. One of these tools is molecular dynamics simulation, which can provide details of the molecular motions of proteins as a function of time.[72] They use simple approximations based on classical mechanics, or more specifically, on Newton's second equation of motion to simulate atomic motions instead of using the complex and computational intensive quantum-mechanical level of calculations. With the reduced computational complexity, molecular dynamics is able to simulate relatively large protein systems.

The force on each atom is determined by the potential energy of the system, U, which is the sum of interaction energies, including non-bonded and bonded interaction energies. The non-bonded interaction energies are modeled by van der Waals energy and electrostatic energy.[73] Van der Waals energy is represented by a combination of the

dispersion and repulsion energies and is often modeled by the Lennard-Jones potential as written in the following form:

$$U(r_{AB}) = 4\varepsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right]$$

where σ_{AB} is the interatomic separation where repulsive and attractive forces balance, r_{AB} is the distance between atom A and atom B, and ε_{AB} is the Lennard-Jones well depth. Both are constants specific to atoms A and B. The electrostatic interaction is represented by the Coulomb potential. Each atom is assigned a partial charge, and the interaction energy between atom A and atom B is

$$U_{AB} = \frac{q_A q_B}{\varepsilon_{AB} r_{AB}}$$

where ε_{AB} is the effective dielectric constant for the medium, r_{AB} is the distance between atom A and atom B, q_A is the partial charge of atom A, and q_B is the partial charge of atom B.

The bonded interaction energies include the bond stretching, valence angle bending, and torsion energies. The bond stretching energies are approximated using a harmonic potential as in the following form:

$$U\left(r_{AB}\right) = k_{AB}(r_{AB} - r_{AB,eq})^2$$

where k_{AB} is the bond force constant, which determines the strength of the bond, and $r_{AB,\,eq}$ is the equilibrium bond length. Both k_{AB} and $r_{AB,\,eq}$ depend on the types of involved atoms. The valence angle bending energy is also represented by a harmonic potential:

$$U\left(\theta_{ABC}\right) = k_{ABC}(\theta_{ABC} - \theta_{ABC,eq})^2$$

where k_{ABC} is the angle force constant and $\theta_{ABC,\,eq}$ is the angle between three-bonded atoms at equilibrium. The torsional potential energy is associated with the steric barriers between 1,4-pairs of atoms. Since torsion is periodic, the torsion potential energy is periodic and is modeled by the periodic cosine function:

$$U(\omega_{ABCD}) = k_{ABCD}\left[1 + \cos(n\omega_{ABCD} - \delta)\right]$$

where k_{ABCD} is the dihedral force constant, n is the multiplicity of the function, ω_{ABCD} is the dihedral angle, and δ is the phase shift. Additional potential energies may include the improper dihedral term, which accounts for out of plane bending energy, and the Urey-Bradley component, which describes angle bending energy using 1,3-nonbonded interactions. These energy terms are parameterized based on sophisticated quantum-mechanical calculations or experimental data in order to represent the behavior of molecules. The parameters and the forms of these energy terms are called “force fields”.

Currently, MD simulations face two major challenges including the limitations of the available force fields and high computational cost. Whenever quantum effects are important, MD simulations cannot accurately predict molecular motions. MD methods cannot handle the situation of bond breaking and formation and the electronic polarization is generally ignored. Although the computational cost of MD has been significantly reduced compared to the quantum-mechanical calculations, simulations longer than a microsecond are still time consuming and are not conducted routinely. Lack of enough simulation time may lead to inadequate conformational sampling and thus key events or key conformational state may not be identified.

Although there are several limitations of MD to overcome, MD simulations have been contributing to the drug discovery process. MD simulations have played an important role in identifying cryptic and druggable allosteric binding sites and in predicting free energy of ligand binding.[74, 75] MD simulations have also improved the identification of true small-molecule binders of enzymes by taking into account the receptor flexibility.[76] Instead of docking compounds into one X-ray crystal or NMR structure, each compound is docked into multiple conformations of the target protein generated by MD simulations. This virtual screening protocol better represents the actual dynamic ligand binding process and therefore reduces the possibility of missing true ligands that are often discarded when using the single experimental structure for virtual screening.

In this chapter, we reported the results and analysis of the MD simulations on Rtt109. The purpose of this MD study is to investigate the flexibility of Rtt109 lysine-binding tunnel. In addition, multiple conformations of Rtt109 generated by our MD simulations may serve as the structures for ensemble docking studies if needed in the future.

3.2 Experimental Section

We performed the MD simulations using the NAMD[77] package with the CHARMM[78, 79] force field for the structure of Rtt109 (2ZFN.pdb[31]) with explicit water. 2ZFN.pdb[31] was chosen because it was the highest resolution structure available at the time the MD experiment was run. Crystallographic water was kept and a truncated

octahedral box of TIP3P water was loaded into the system to form the protein environment. The system was first energy minimized for 500 conjugate gradient steps. The water molecules were then relaxed through a 100 ps simulation with constant temperature and pressure and then a 400 ps simulation with constant temperature. The temperature and pressure were controlled and maintained at 310 K and 1 atm using the Langevin thermostat and Langevin piston barostat control methods. Subsequently the whole system was equilibrated for 500 ps under NpT (constant temperature T, pressure p, and number of particles N) conditions and 1 ns under NVT (constant temperature T, volume V, and number of particles N) conditions. A 10-ns simulation was then performed under a constant temperature of 310 K.

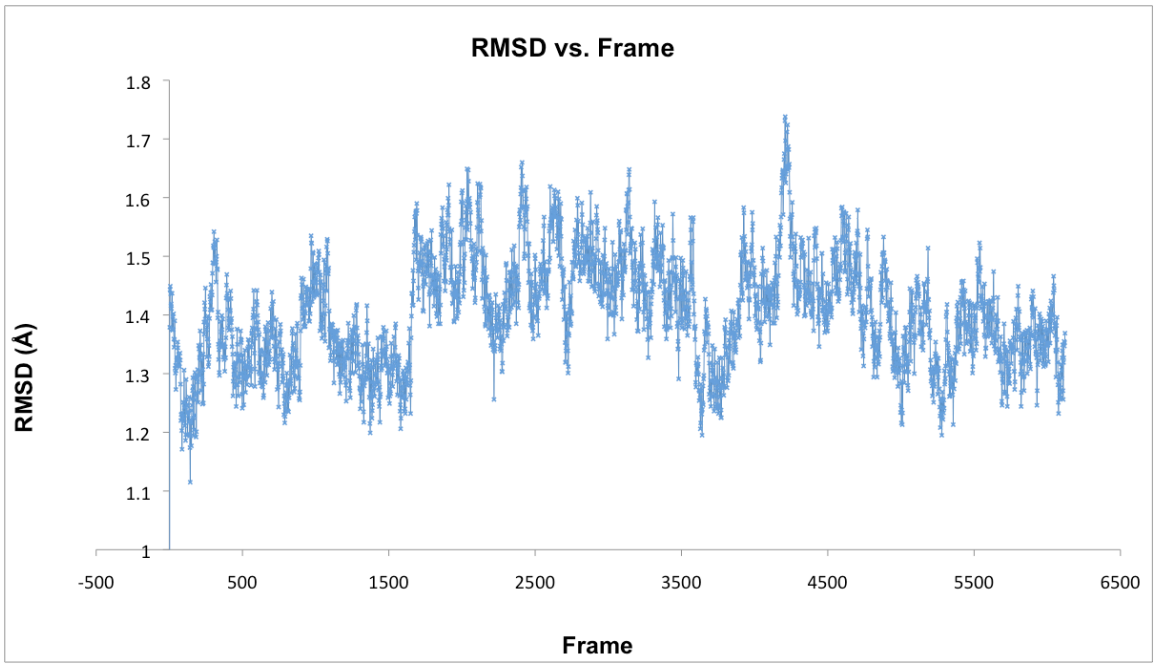
3.3 Results and Discussion

Protein flexibility was assessed by calculating the root mean square deviation (RMSD) between the structures generated every 2 ps from the MD simulations and the crystal structure. A comparison was then made with the experimental data as a means of validation.

All residues in the simulation were aligned by rigid-body translations and rotations and compared to the crystal structure of Rtt109. The average RMSD value of residues 1 to 124 and 175 to 404 backbone atoms (excluding missing residues and tail residues) compared to the crystal structure of Rtt109 during the 10 ns simulation is 1.40 Å (Figure 3.1), while the average RMSD of all residue backbone atoms is 3.65 Å. The difference between the RMSD values indicates that the flexibility of the tail residues

(residues 125 to 142) contributes largely to the RMSD value of all residue backbone atoms. The overall average RMSD value of all residue atoms excluding hydrogen atoms compared to the starting structure for the 10-ns simulation is 1.81 Å. The RMSD was generally stable, indicating the protein was not very flexible. Within the 10-ns simulation, no sharp increases in RMSD suggested that no significant conformational changes occur.

Figure 3.1. The RMSD value of residues 1 to 124 and 175 to 404 backbone atoms compared to the crystal structure of Rtt109 during the 10 ns simulation



B-factors in crystallography are assumed to indicate the relative dynamics of different parts of the structure. Atoms with low B-factors generally are parts of the structure that are well-ordered. Atoms with large B-factors usually indicate that the corresponding part of the structure is very flexible.[80] In our study, we compared the B-factor of the Cα atom of each residue (Figure 3.2) and the average RMSD value of the Cα atom of each residue (Figure 3.3). We found the general trend for B-factor correlated well with that for the calculated RMSD values during the 10-ns MD simulation. As shown in Figure 3.2 and Figure 3.3, the residues with peak B-factor values usually have peak RMSD values. These residues include residues 40-41, 76, 95-97, 130-132, 175, 251, 273-274, 299-301, 321-322, 353-355, and 389-391, which have both peak values or close to peak values of B factor and RMSD. These residues are often located at the surface of the protein as shown in Figure 3.4.

The most important area of Rtt109 for our study is the lysine/AcCoA-binding tunnel. The residues 65-68, 82-89, 189-200, 213, 215, 218-219, 222, 226, 285-293, and 323-330 in the lysine-AcCoA binding tunnel were used in our docking studies to represent the active site. The average RMSD value of these residues during the 10-ns MD simulation compared to the crystal structure was 1.78 Å, which indicated the active site was not very flexible. The five residues with the highest RMSD values among the residues in the lysine/AcCoA-binding site were Pro286, Asp287, Arg292, Leu323, and Ser324, which are listed in Table 3.1 and displayed in Figure 3.5. It is obvious from Figure 3.5 that these five residues are also located at the surface of the protein where residues are usually more flexible than those that are buried inside the protein.

Figure 3.2. The B-factor values for each residue in Rtt109

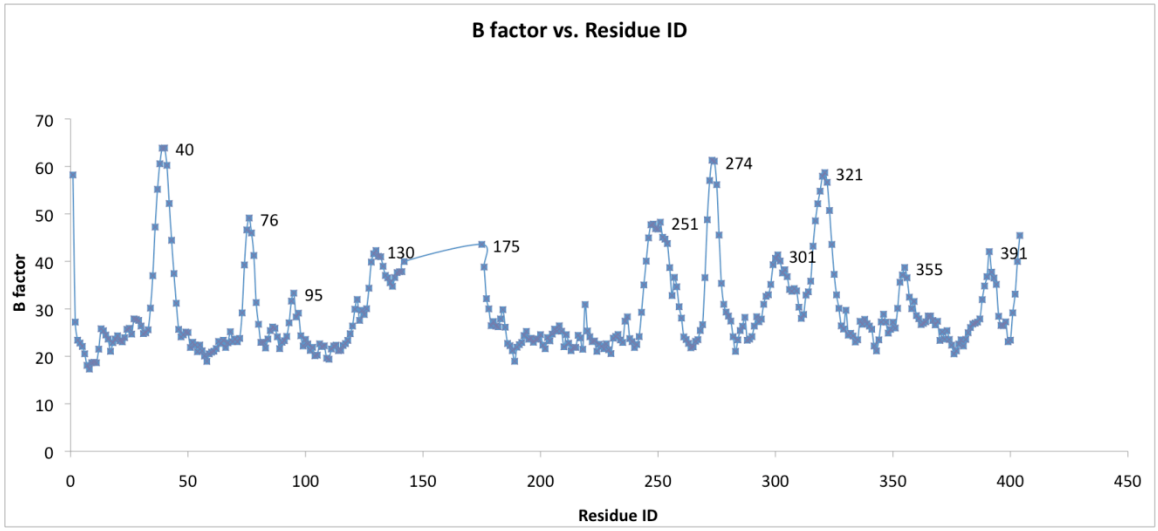


Figure 3.3. The average RMSD of Cα atom of each residue during the 10-ns simulation

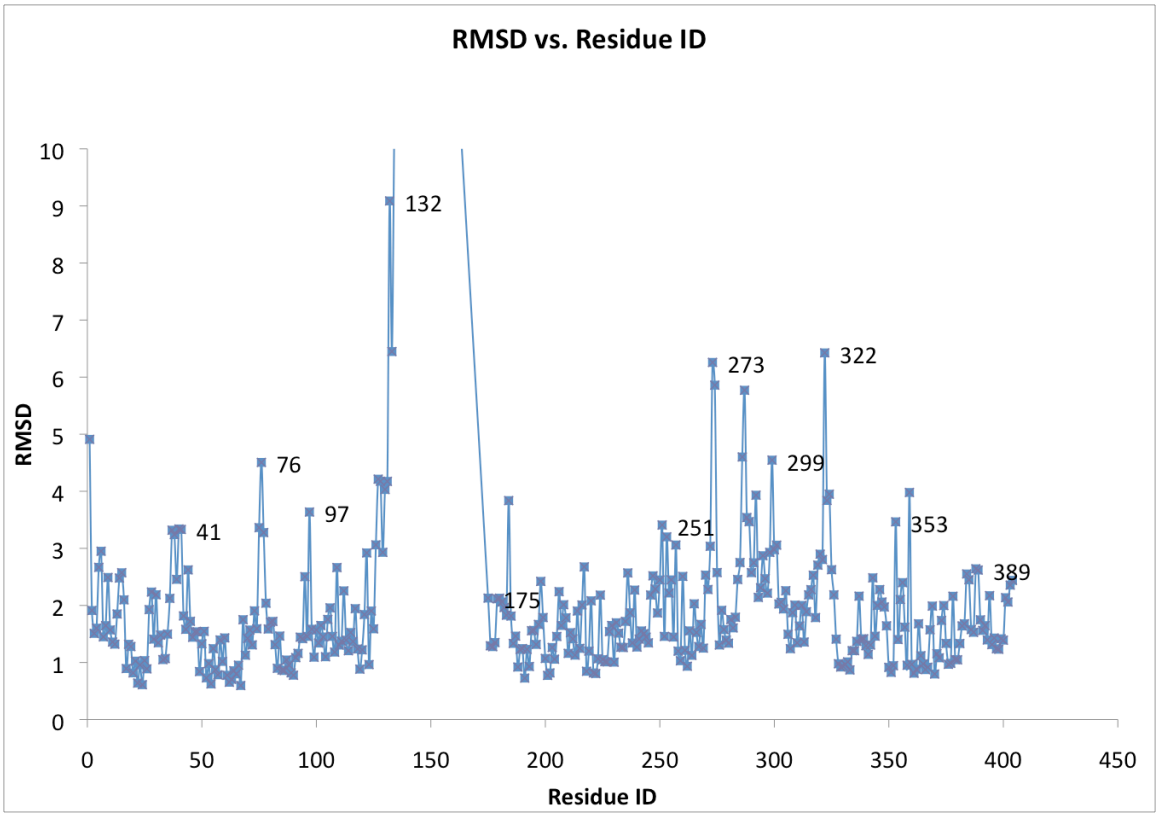
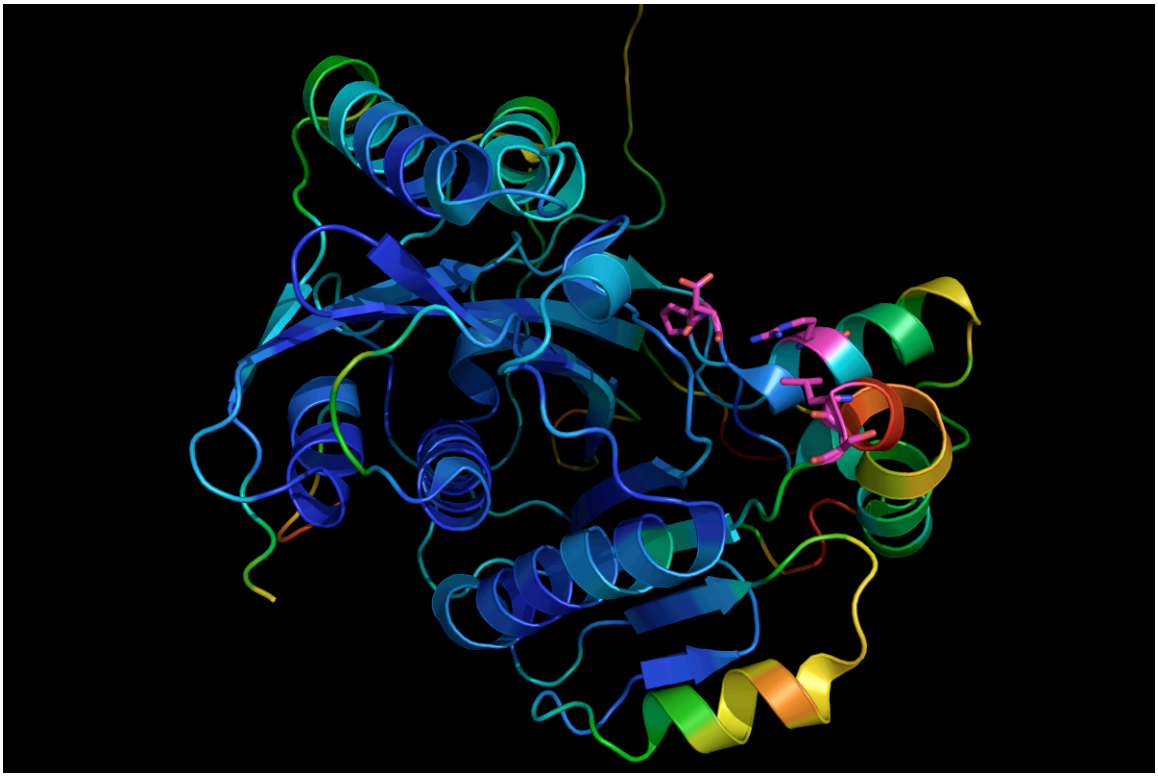


Figure 3.4. Residues (magenta) with both peak values or close to peak values of B-factor and RMSD

Table 3.1. The five residues with the highest RMSD values among the residues in the lysine/AcCoA-binding site

Residue ID	RMSD (Å)
Asp287	5.77
Pro286	4.60
Ser324	3.95
Arg292	3.93
Leu323	3.84

Figure 3.5. Five residues (magenta) with the highest RMSD values among the residues in the lysine/AcCoA-binding site



3.4 Concluding Remarks

A 10-ns MD simulation was performed on Rtt109 with explicit water at 310K. No significant conformational changes were identified during the 10-ns MD simulation process. For the 10-ns simulation, the overall average RMSD value of all residue atoms excluding hydrogen atoms compared to the equilibrium structure is 1.81 Å. The protein dynamics based on the MD simulation showed a trend similar to the experimental B-values in the X-ray crystallography. Residues 40-41, 76, 95-97, 130-132, 175, 251, 273-274, 299-301, 321-322, 353-355, and 389-391 have both peak values or close to peak values of B factor and RMSD. The most flexible residues reside at the surface of the protein, including Pro286, Asp 287, Arg292, Leu323, and Ser324. Overall, the whole protein and the lysine/AcCoA-binding site are not very flexible. These results indicate that the use of one conformation, the crystal structure of Rtt109, 2ZFN.pdb, for docking studies should be appropriate and should not produce large numbers of false negatives.

CHAPTER 4:

VIRTUAL SCREENING STUDIES USING DOCKING AND SCORING

4.1 Introduction

Docking and scoring is a process that places a molecule (ligand) in the proposed binding site of its macromolecular target (receptor) and estimates its binding propensity. A successful docking and scoring should be able to accurately (compared to the experimental results) predict the binding pose (conformation and orientation) of a ligand and its binding affinity.[81] Docking and scoring has been widely used in the discovery and development of novel drugs. Many successful cases, including the identification of HIV protease inhibitors, have shown that the structure-based design via docking and scoring heavily influenced the development of drugs.[82] When only the target structure is available, virtual screening of compound database by docking and scoring usually serves as the most applicable computational tool for hit identification. Docking and scoring can also be utilized during the lead optimization process. Modifications to known active compounds can be prioritized based on the binding affinity predicted by docking and scoring before synthesis.

To identify novel candidate scaffolds for Rtt109 inhibition as well as crucial ligand/substrate and receptor interactions, we have used Surflex-Dock[23, 83, 84] (Tripos, Inc.) and Glide[85-87] (Schrödinger, Inc.) to dock small molecule libraries and histone substrate mimics. Glide was mainly used to dock the histone substrate mimics into the

proper position of Rtt109, because Glide has the functionality to dock compounds with user-defined constraints that facilitates the docking of the flexible peptides.

4.1.1 Surflex-Dock

The Surflex-Dock utilizes the fragmentation and similarity search techniques.[84] An idealized active site ligand (protomol), which complements the receptor active site, is first generated as a template for computing putative binding poses of a molecule or a molecular fragment. To generate the protomol, three types of molecular probes are placed into the receptor active site. The molecular probes include CH_4 as a steric (hydrophobic) probe, C=O as a hydrogen-bond acceptor probe, and N-H as a hydrogen-bond donor probe. Each probe's position and orientation are optimized based on the scoring function, in order to maximize its interaction with all protein atoms.[88] Low-scoring, isolated, and redundant probes are eliminated. The remaining probes collectively constitute the protomol.

The second step involves the molecular fragmentation.[84] To reduce the computational time in conformational sampling, each molecule is fragmented by breaking its non-ring rotatable bonds. To achieve molecular flexibility of ligands in docking, a heuristic set of rules are applied to the conformational sampling, where two, three, or six rotamers are used for each bond. In addition, the algorithm picks the most different conformations per fragment based on the root-mean-square deviation. The user can specify the maximum number of conformations per fragment.

The third step is to align the fragments and construct the full molecules.[84] All conformations of all molecular fragments are aligned with the "protomol" based on

similarity. The aligned fragments are scored and pruned according to the scoring function. Then using a "Hammerhead" approach, the conformations of full molecules are constructed from the aligned fragments.[89] The best scoring poses (conformations) are optimized and returned with their scores.

Surflex-Dock uses an empirically derived scoring function that is based on the binding affinities of experimentally determined protein-ligand complexes, with its output scores expressed in $-\log_{10}(Kd)$ units as the estimated binding affinities.[90] The scoring function is a weighted sum of non-linear functions, including the following terms:

1) Hydrophobic: a weighted sum of a Gaussian-like function and a sigmoid function over all atom pairs, which captures the positive atomic contact and the negative atomic contact due to atomic interpenetration.

2) Polar: a sum over all pairs of complementary polar atoms to represent the effects of hydrogen bonds and salt bridges.

3) Entropic: a penalty term that is linearly related to the number of rotatable bonds in the ligand and to the log of the molecular weight of the ligand, intending to model the loss of translational and rotational entropy of the ligand once the ligand is docked.

4) Solvation: a term that is linearly related to a count of the difference between the potential and actual numbers of hydrogen bond equivalents.

4.1.2 Glide

In Glide, the shape and other properties of the receptor are computed and represented on a grid by binned site points and receptor distances before the docking process.[87] In the next step, an exhaustive conformational and position search of the ligand

over the active site of the receptor is conducted, followed by a heuristic screen that eliminates high-energy conformations and other unsuitable binding conformations. The remaining conformations of the ligand are minimized within the receptor using a molecular mechanics energy function of the OPLS-AA force field.[91] After that, three to six lowest-energy poses are subjected to a Monte Carlo process that explores nearby torsional minima. An extended and modified version of the ChemScore[92] funcion, GlideScore is utilized to predict binding affinity and rank ligands in the database virtual screens. The GlideScore function includes a lipophilic-lipophilic term, three differently weighted hydrogen-bonding terms, a metal-ligand interaction term, a polar-hydrophobic term, terms for the Coulomb and vdW interaction contributions, and terms to deal with the solvation effects. However, Glide uses the Emodel function which has a more significant weighting of electrostatic and van der Waals energies to select the docked pose.

4.2 Experimental Section

To identify potential scaffolds for inhibitor design, I screened approximately 7 million non-redundant compounds *in silico* for potential activity against Rtt109. These structures were obtained from five small-molecule databases: DrugBank[93], LeadQuest[94], the University of Minnesota in-house "Gopher" (GPHR) library, the University of Minnesota Institute for Therapeutics Discovery and Development (ITDD) compound database, and a novel peptoid (poly-*N*-substituted glycines) probe library generated by the Walters laboratory. All compound databases were first refined using a set of three

filters in the SciTegic PipelinePilot data analysis and reporting platform (Accelrys, Inc.)[95], retaining only those structures that satisfied Lipinski's Rule of Five[96], and did not contain inorganic atoms and/or reactive substructures. Predicted bound configurations for these structures were obtained using Surflex-Dock (SYBYL 8.0, Tripos, Inc.), with 2ZFN.pdb[31] representing the Rtt109/AcCoA binary structure. The protomol (active-site representation) corresponded to residues 65-68, 82-89, 189-200, 213, 215, 218-219, 222, 226, 285-293, and 323-330 in the Lys-AcCoA binding tunnel. Docked poses were ranked by total Surflex-Dock score expressed as $-\log(K_d)$. Threshold and bloat parameters were set to 0.5 and 0, respectively. The maximum number of conformations per compound fragment and the maximum number of poses per molecule were both set to twenty, and the maximum allowable number of rotatable bonds per structure was limited to 100. Post-dock minimizations were carried out on each ligand to optimize predicted configurations in the receptor site.

To identify potential substrate and receptor interactions, several peptide pieces were docked into the Rtt109 lysine binding area using Glide. These peptides contained the lysine residue that was acetylated by Rtt109 and its neighbor residues on the histone tail to mimic the substrate. Using 2ZFN.pdb[31] with AcCoA removed, a Glide grid was computed with an outer box region of 20 Å to represent the receptor. The grid midpoint was defined by the centroid of the lysine entrance tunnel that was composed of residues 65-68, 82-89, 189-200, 213, 215, 218-219, 222, 226, 285-293, and 323-330.

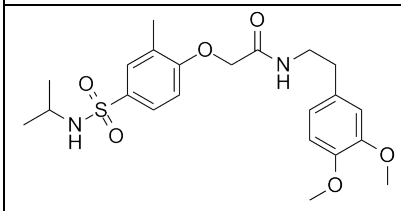
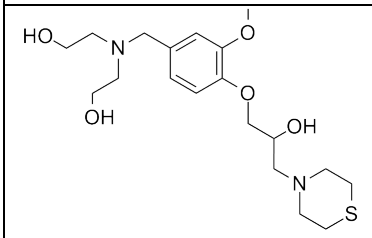
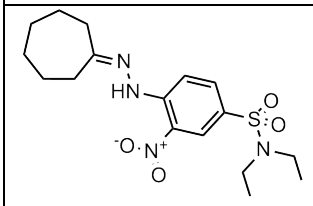
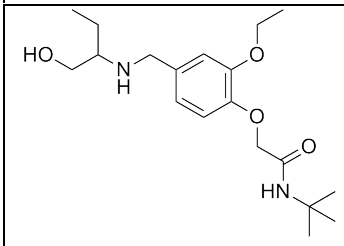
4.3 Results and Discussion

Compounds were retained which exhibited a Surflex-Dock score of 10 or greater, corresponding to predicted K_d values of 0.1 nM or lower, resulting in 961 potential hits. This virtual screening process has yielded 6 DrugBank, 32 LeadQuest, 83 GPHR, 747 ITDD, and 93 peptoid compounds (Table 4.1). The five highest-scoring compounds from the overall screening process are shown in Table 4.2 together with corresponding Surflex-Dock scores. Figure 4.1 illustrates the docked configurations of these compounds along with key ligand-receptor interactions predicted by the docking. These compounds share some common features: a linear scaffold which fits well with the narrow lysine binding tunnel and several polar functional groups on the scaffold which make H-bonding or Coulombic interactions with residues inside the lysine tunnel. The highest scoring compound, **4.1** (Total Score = 12.43), was predicted to donate a H-bond to the backbone CO of Asp287 as shown in 2D and 3D interaction maps (Figure 4.2 and Figure 4.3). Its propyl group fits within the small hydrophobic area formed between Phe84 and Leu191. The oxygen atoms on the sulfonamide group point to the solvent-exposed area and make favorable Coulombic interactions with the polar residues at the entrance of the lysine tunnel. The phenyl ring can π-π interact with the Tyr199 side chain while having hydrophobic interactions with other nearby hydrophobic residues. The Trp222 side chain and the Leu213 backbone were predicted to donate a H-bond to the CO and ether O atoms of compound 4.1, respectively. Favorable hydrophobic interactions can be formed between the 1,2-dimethoxybenzyl group of compound 4.1 and the residues Phe62, Ile212, and Leu213.

Table 4.1. Virtual screening results

database	# of cpds in database	# of cpds scoring 10 or higher	preliminary hit rate (%)
DrugBank	4406	6	0.14
LeadQuest	53640	32	0.06
GPHR	208593	83	0.04
ITDD	1050000	747	0.07
Peptoids	11046	93	0.84

Table 4.2. Five highest-scoring compound structures from among all compounds screened to date, with total Surflex-Dock scores

Structures	Compound ID	Total Score
	4.1	12.43
	4.2	12.41
	4.3	12.32
	4.4	12.28

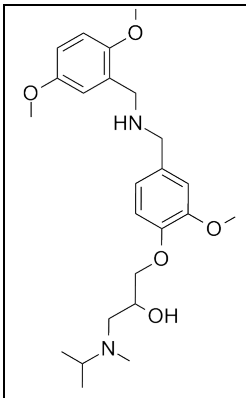
	4.5	12.25

Figure 4.1. Three-dimensional renderings of five highest-scoring compounds (hydrogens undisplayed) docked into Rtt109 (2ZFN.pdb[31]) (SYBYL 8.0, Tripos, Inc.)

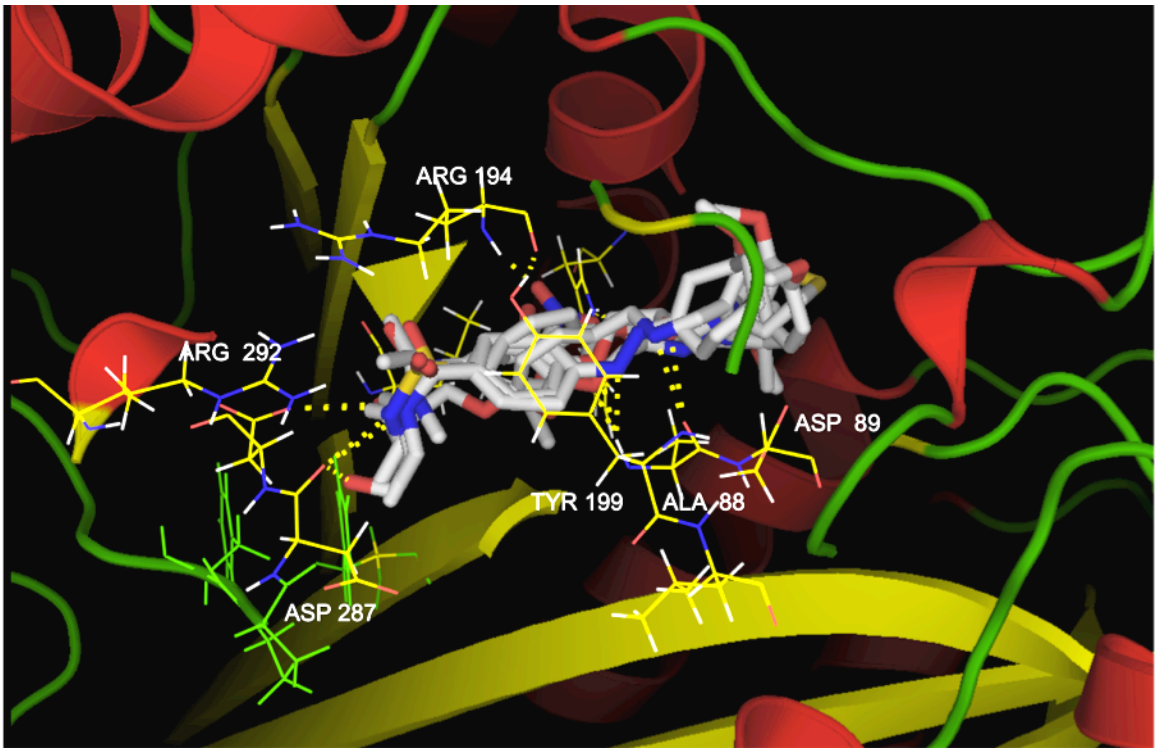


Figure 4.2. Two-dimensional interaction map of compound 4.1 with Rtt109 (MOE 2010.10, Chemical Computing Group, Inc.)

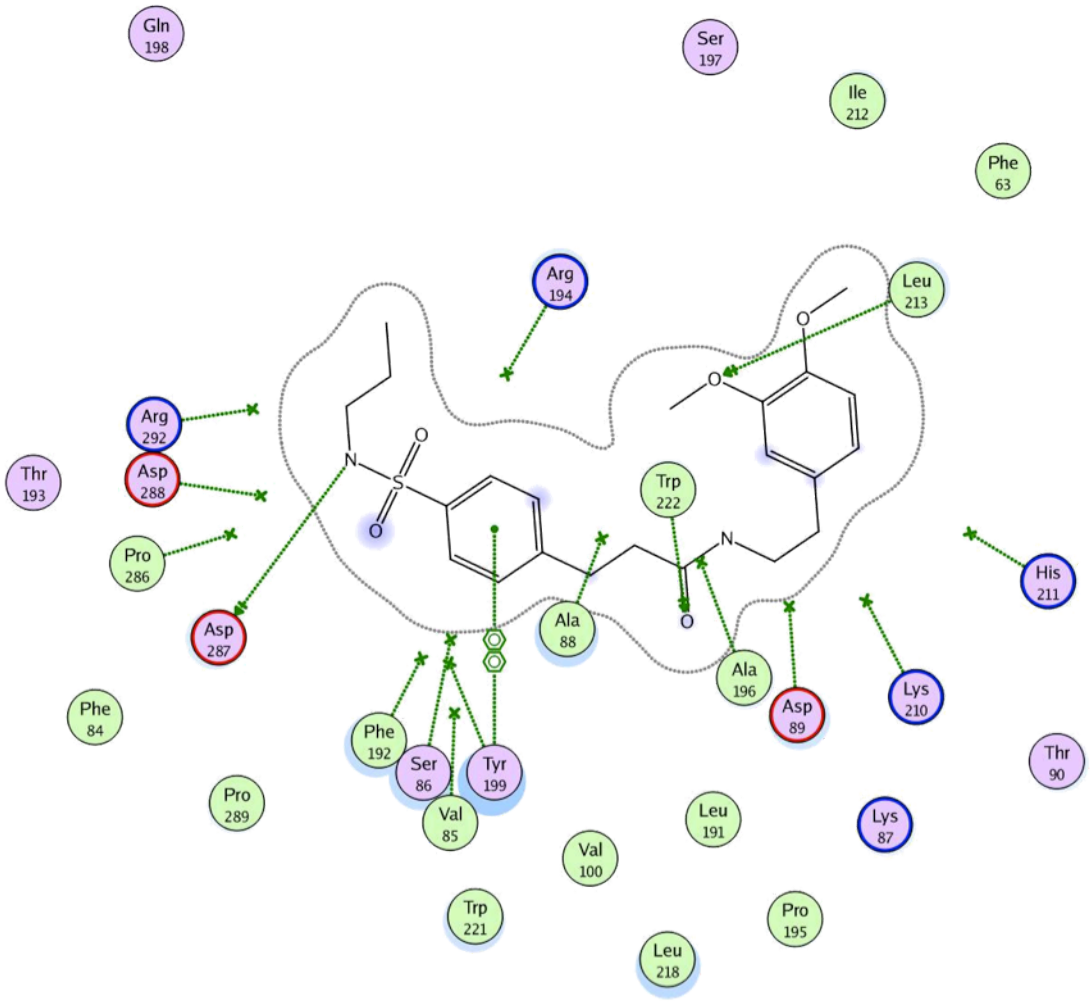
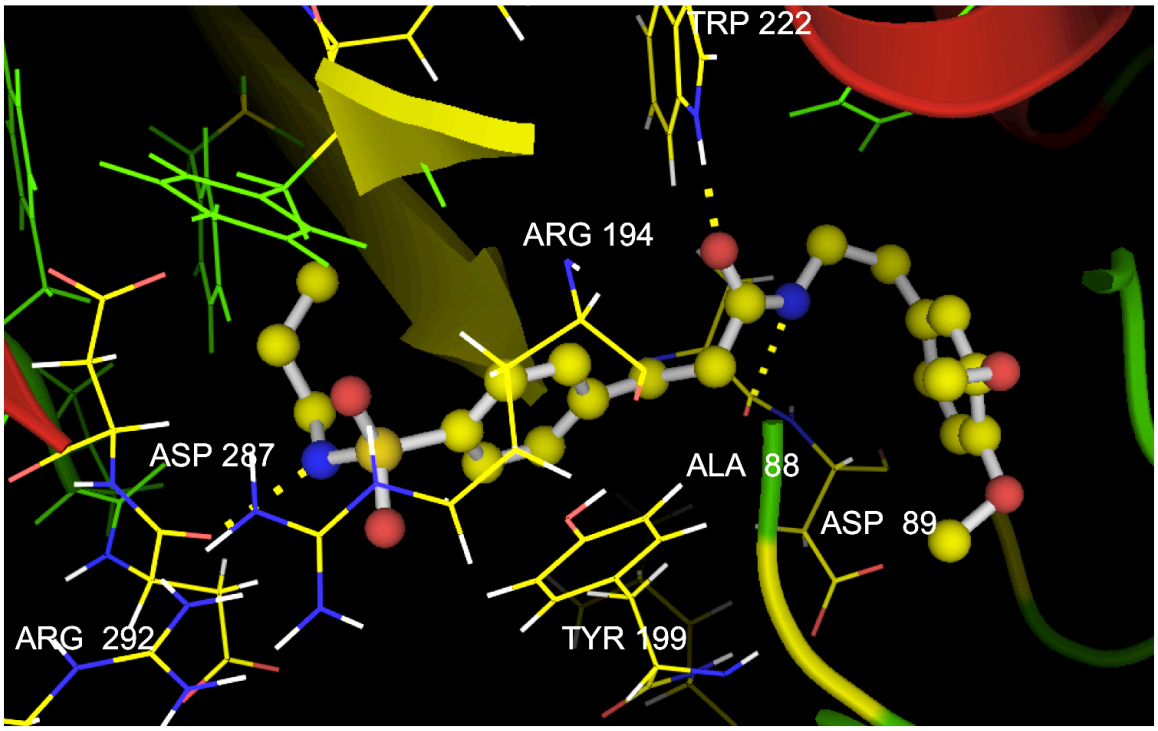


Figure 4.3. Three-dimensional interaction map of compound 4.1 with Rtt109 (SYBYL 8.0, Tripos, Inc.)



To better understand the potentially important interactions between compounds and the lysine tunnel, we analyzed the residues that were frequently involved in the interactions between the protein and the top scoring compounds from different databases in our docking studies. These residues include Pro286, Ser86, Arg292, Arg194, Asp288, Asp287, Trp222, Val85, Ala88, Phe192, Lys87, Ala196, and Tyr199. The details of the frequency and the type of interactions of these residues are shown in Table 4.3. Interestingly, single-site mutagenesis studies have shown that residues Tyr199, Trp222, Asp287 and Asp288 are important for the catalysis and/or substrate binding of Rtt109 and these four residues are also predicted to be key residues that interact with the compounds in the lysine-binding tunnel.[31, 38, 44] Thus these residues may be most important residues to target when designing Rtt109 inhibitors.

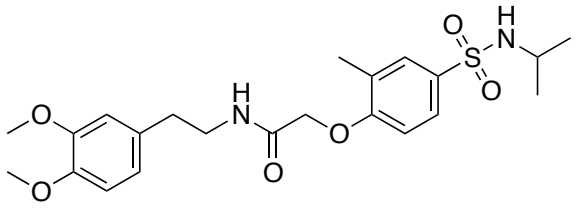
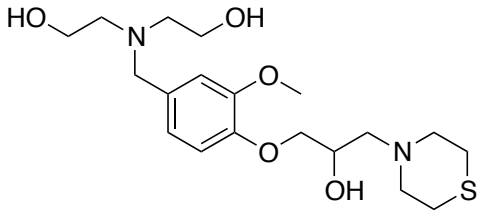
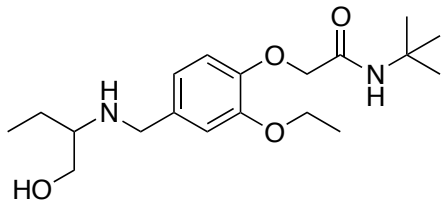
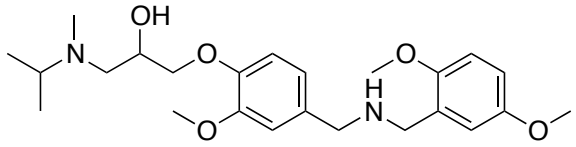
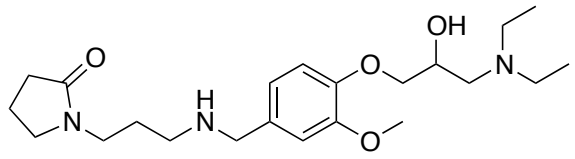
Twenty-two commercially available compounds (see Table 4.4) with highest scores were experimentally examined for their inhibitory activities against Rtt109 using the HTS2 method described in Chapter 2. None of them showed activity at the concentration of 250 μM. It is possible that minor side chain movements in the narrow active site may impact on the prediction of affinity and pose of compounds. Although molecular dynamics simulations can be applied to capture some of these movements, it is very difficult to identify the important but subtle side chain movements that will actually contribute to compound binding from millions of the snapshots generated by the molecular dynamics simulations. In addition, very little experimental data were available to allow us to conduct the benchmark work. It is also possible that the sampling of ligand conformations did not capture the ligand-binding conformation among all docking poses.

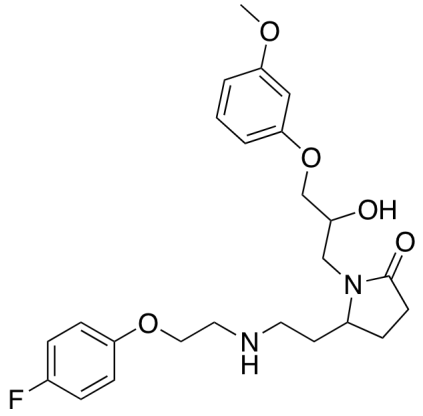
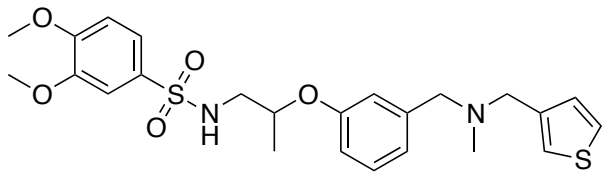
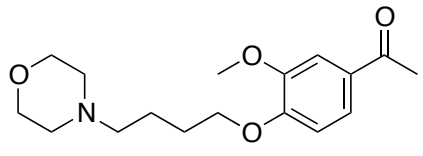
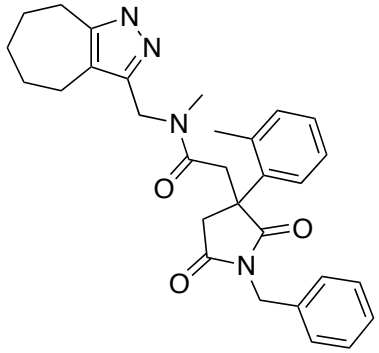
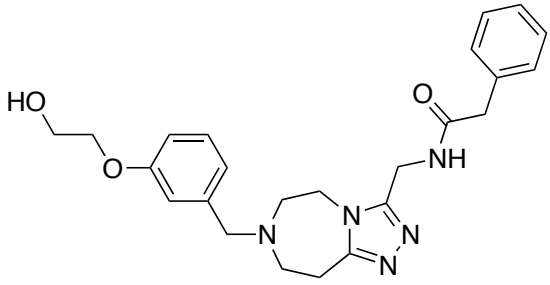
The scoring function was also not guaranteed to linearly rank compounds according to their inhibitory activities. The solvation energy and the associated entropy penalty for ligand binding may not be taken fully into account in the scoring functions. Therefore, for our highest scoring compounds that contain many polar groups, this inaccurate estimation may lead to relatively larger errors compared to those of compounds that contain fewer polar groups.

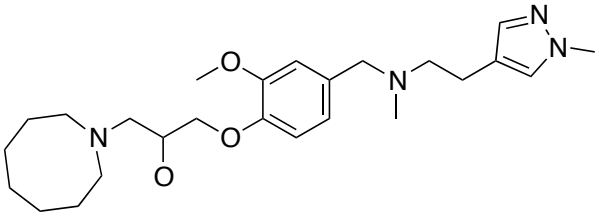
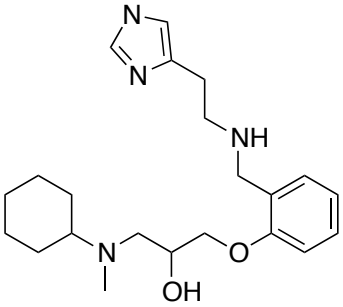
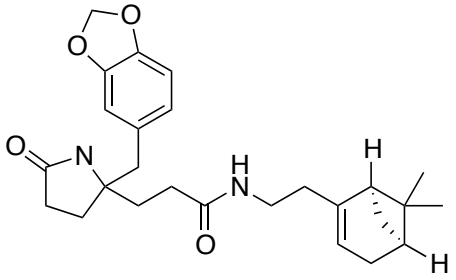
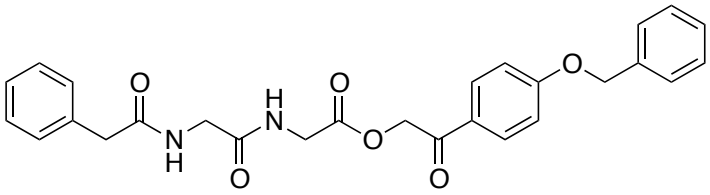
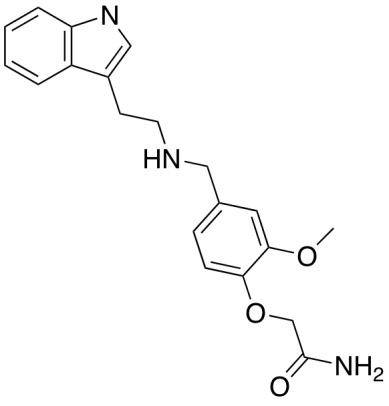
Table 4.3. Key residues that are predicted to interact with the ligands (A: H-bonding acceptor; D: H-bonding donor; π+: π-cation interactions; π π: π-π interactions)

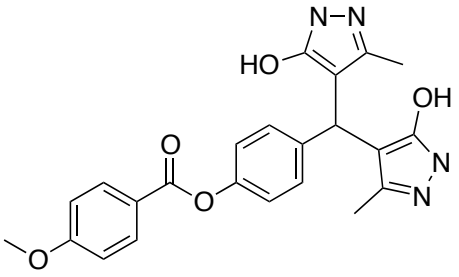
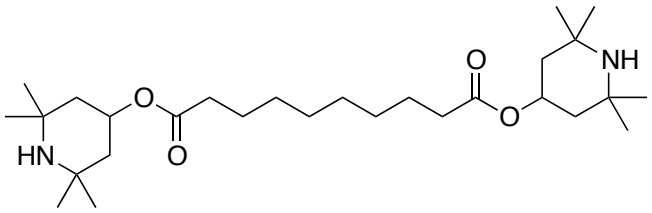
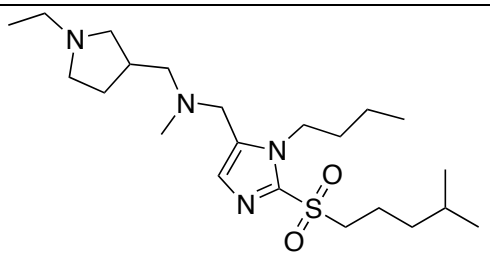
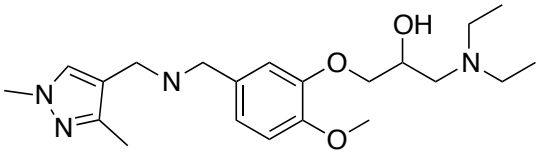
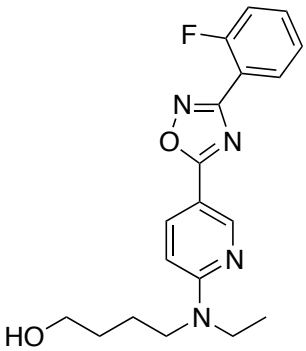
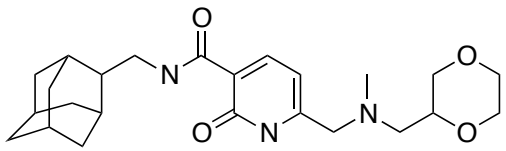
Residues	Frequency Count (top 22 ligands from peptoids)	Frequency Count (top 30 ligands from GPHR)	Frequency Count (top 12 from ITDD vendor database)
Pro286 (A)	16	2	0
Ser86 (A)	15	1	0
Arg292 (D)	13	5	5
Arg194 (D)	13	6	6
Asp288 (A)	12	3	1
Asp287(A)	10	9	3
Trp222 (D)/(π+)	10	15	5
Val85 (A)	7	2	0
Ala88 (A)/(D)	7	22	6
Phe192 (A)	5	3	2
Lys87 (π+)/(D)	5	1	0
Ala196 (D)	4	3	1
Tyr199 (π π)	1	23	5

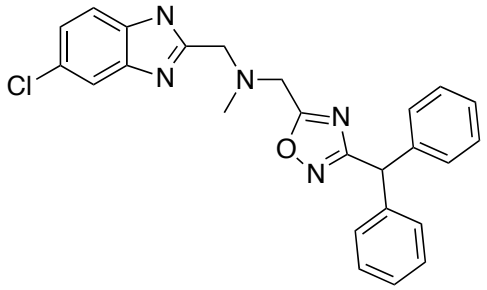
Table 4.4. Twenty-two commercially available compounds that were experimentally examined for their inhibitory activities against Rtt109 using the HTS2 method described in Chapter 2

Structures	Total Score
	12.43
	12.41
	12.28
	12.25
	12.18

	11.39
	11.33
	11.16
	11.07
	11.01

	10.93
	10.81
	10.80
	10.79
	10.78

	10.47
	10.44
	10.40
	10.37
	10.08
	10.05

	10.01

In the docking studies of the peptide substrate mimics, the tripeptide Q-K-S with N-terminal acetyl and C-terminal amide capping groups was first docked into the Rtt109 lysine entrance tunnel. The lysine tail of the tripeptide goes into the lysine entrance tunnel as the Rtt109 mechanism suggests and has a hydrogen-bonding interaction with Ser86 of Rtt109. The Gln and Ser moieties of the tripeptide is proposed to interact with residues Tyr199, Lys87, Asp287, Gln198, and Arg194 near the surface or on the surface of Rtt109 as shown in Figure 4.4. A pentapeptide Y-Q-K-S-T was then docked using the same method with one constraint, that the lysine side chain of the pentapeptide and Rtt109 residue Ser86 need to form a hydrogen bond. This constraint was added in order to make sure the lysine portion of the pentapeptide dock into the lysine entrance tunnel instead of the surface of Rtt109. Figures 4.5 and 4.6 illustrate the detailed interactions between the pentapeptide and Rtt109. Residues Ser86, Asp287, Lys87, and Arg292 interact with the pentapeptide. These residues may play important roles for the histone binding and guide the design of inhibitors that block the protein-protein interaction between Rtt109 and histone.

Figure 4.4. Two-dimensional interaction map of the tripeptide with Rtt109 (MOE 2010.10, Chemical Computing Group, Inc.)

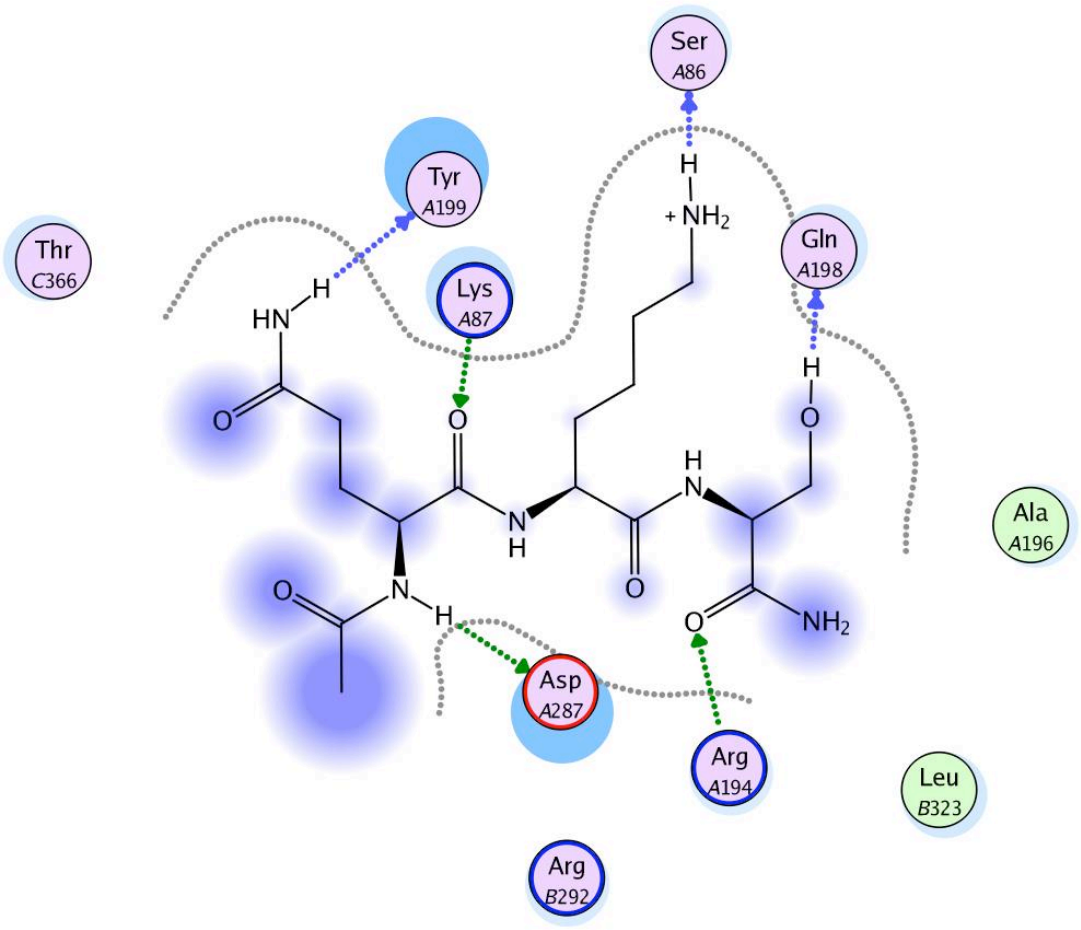


Figure 4.5. Two-dimensional interaction map of the pentapeptide with Rtt109 (MOE 2010.10, Chemical Computing Group, Inc.)

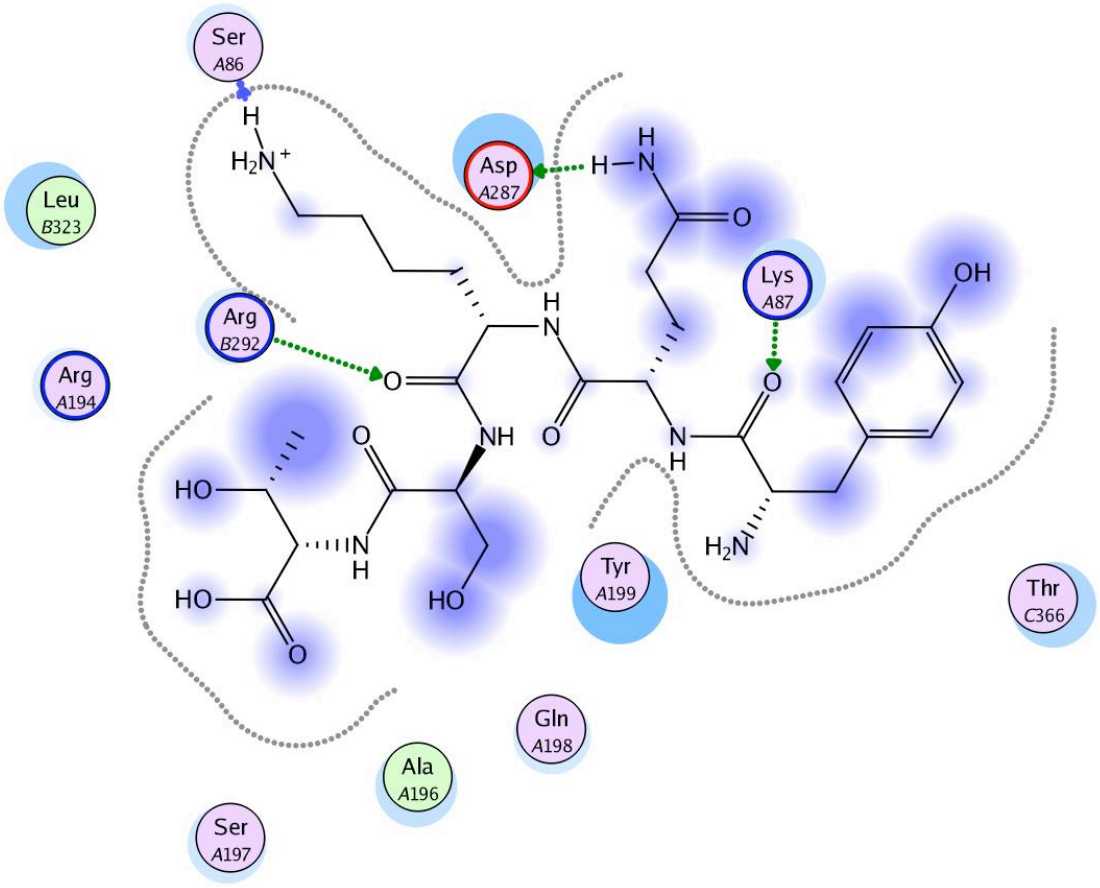
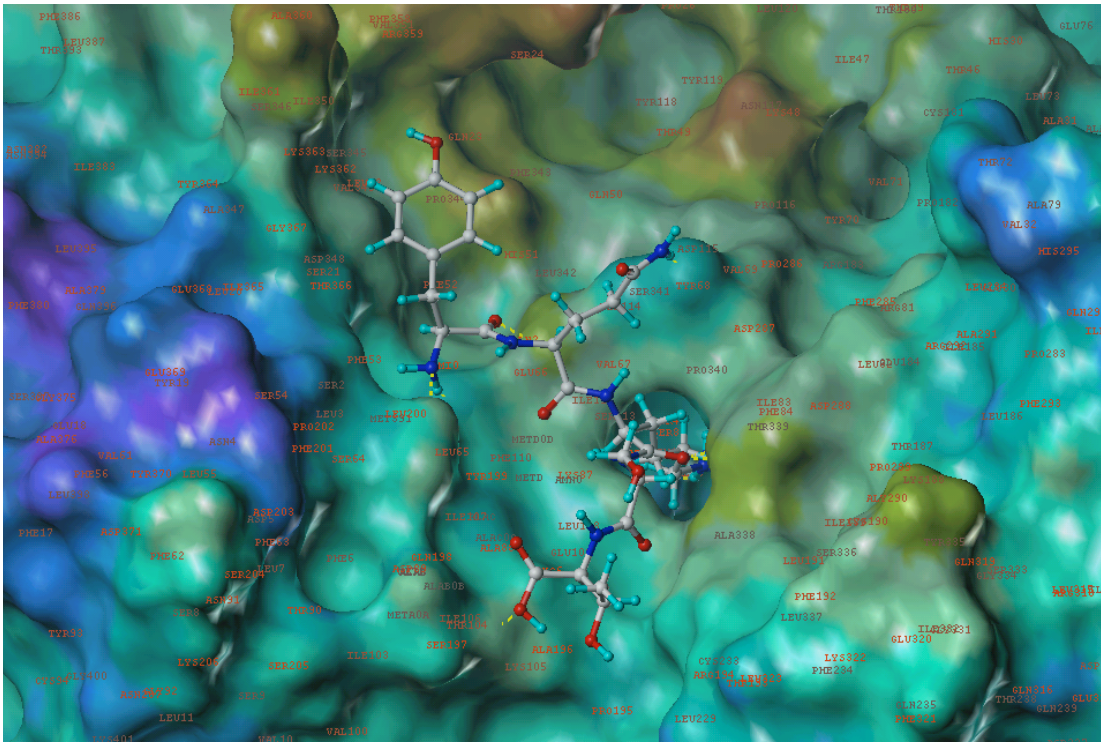


Figure 4.6. Three-dimensional interaction map of the pentapeptide with Rtt109 (SYBYL 8.0, Tripos, Inc.)



4.4 Concluding Remarks

Understanding the structural bases for molecular recognition between an enzyme protein and its inhibitors is important for the identification of new inhibitors and the optimization of known inhibitors. Structural biology experiments such as X-ray crystallographic and NMR studies play key roles in helping researchers understand these structural bases. Docking and scoring studies take advantage of the structural information and aid the rational drug design. In our studies, we utilized docking and scoring as a virtual screening tool in order to identify novel inhibitors against Rtt109. By analyzing the predicted key ligand-receptor interactions, we found that residues Pro286, Ser86, Arg292, Arg194, Asp288, Asp287, Trp222, Val85, Ala88, Phe192, Lys87, Ala196, and Tyr199 may be the key residues that contribute to the ligand binding. Among them, residues Tyr199, Trp222, Asp287 and Asp288 have been demonstrated experimentally to be important for the catalysis and/or substrate binding of Rtt109 and thus they should be the most important residues to target for designing Rtt109 inhibitors. We also explored the potential interactions between the substrate histone and Rtt109 on the surface between these two proteins by docking truncated peptide fragments of histone. Ser86, Asp287, Lys87, and Arg292 of Rtt109 are predicted to be major residues that interact with histone. These residues may also need to be targeted by Rtt109 inhibitors in order to efficiently interrupt the interactions between Rtt109 and histone. Twenty-two highest scoring compounds were tested against Rtt109 using the HTS2 assay method but demonstrated negligible activities experimentally. The reasons for these results may be: 1) the fact that subtle side chain movements may make major differences in the structure of a small and

narrow tunnel, and 2) the limitations of the docking and scoring on our protein system. The X-ray crystallographic studies are ongoing. With further information from the structural biology studies, we will refine and improve our docking and scoring studies in the future.

CHAPTER 5:

THREE-DIMENSIONAL QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP STUDIES ON Rtt109 INHIBITORS

5.1 Introduction

Three-dimensional quantitative structure-activity relationship (3D-QSAR) is a method for building a statistical model which correlates biological activities of compounds with their three-dimensional structural properties. Based on the assumption that receptor-ligand interactions are shape- and structure-dependent, 3D-QSAR methods can be employed to elucidate binding characteristics, predict biological activities of novel structurally similar compounds, and improve the potency of known inhibitors.[41, 97] 3D-QSAR is a "ligand-based" technique that requires no initial active-site structural information and is particularly effective in cases when the targeted receptor is unknown or difficult to model, or, in the case of Rtt109 inhibitor design, when the target is known but the active-site location is uncertain.[98]

There are usually several additional assumptions in 3D-QSAR methods, including the following:

1) All compounds in training series interact with the same binding site on the same receptor in the same manner.

2) Interactions that lead to biological effects are non-covalent and mainly enthalpic.

3) The entropic cost upon the binding of compounds in training series follows a similar pattern.

4) The biological activities are caused by the ligands themselves and not by their metabolites or degraded portions.

Thus, in order to build reliable 3D-QSAR models, the selected compounds as the training set need to satisfy the requirements above.

5.1.1 CoMFA and CoMSIA

Comparative Molecular Field Analysis (CoMFA[99]) and Comparative Molecular Similarity Indices Analysis (CoMSIA[100]) are two commonly used 3D-QSAR methods. In CoMFA, all molecules in the training set are aligned using manual or automated methods. Ligand molecules are represented by their steric and electrostatic fields.[99] To calculate the steric and electrostatic fields, the ligand molecules are placed in a three-dimensional lattice grid that is large enough to cover all molecules in this series. A 2.0-Å separation between lattice points is chosen to make the computation efficient. A probe atom with the van der Waals properties of an sp^3 carbon atom and a charge of +1.0 is placed at the intersections of the lattice. The energies of steric interaction between a ligand and a probe atom are calculated using Lennard-Jones 6-12 potential with parameters from the standard Tripos force field.[101] The energies of electrostatic interaction between a ligand and a probe atom are calculated based on the Coulomb potential function, and the atomic charges are calculated using a method of the user's choice. Both the Lennard-Jones 6-12 potential and Coulomb potential curves are very steep when the distance is close to the van der Waals surface. In addition, the Lennard-Jones and Coulomb potentials demonstrate singularities at atomic positions. Thus, cutoff

values (30 kcal/mol by default) are used to avoid unacceptably large values and dramatic changes in surface descriptions. These interaction energy or field values and the ligands' biological activity values constitute a data table with many more columns than rows. The partial least-squares (PLS) technique is then used to correlate the field values with the biological activity data. The PLS analysis resembles the principal component regression analysis and has the advantages of deriving robust linear regression equations from tables with many more columns than rows (more independent variables than samples). A certain percentage (e.g., 5%) of the increase of the cross-validated r_{cv}^2 serves as a criterion to extract only significant components because too many components will lead to an overfitting of the model. The cross-validation evaluates the internal predictivity of a CoMFA model and also determines the optimum number of the PLS components. The PLS regression analysis is then applied without cross-validation using the optimum number of components, resulting in the final models from which conventional correlation coefficient r^2, its standard errors, and *F* ratios (ratios of r^2 to $1-r^2$) are calculated. Comparisons between the predicted and experimental activities are made to identify potential outliers (observation points distant from the majority). The outliers may be removed from the training set to yield a better model with an improved r_{cv}^2, r^2, and *F* ratio. Because of the thousands of coefficients in the regression equation, a direct interpretation of the equation is difficult and not very useful. Instead, the result of the analysis is presented by three-dimensional contour maps. These maps are generated using the products of the coefficients and standard deviation and indicate lattice points or areas where changes in CoMFA field values strongly associate with changes in biological

activities.

As discussed above, there are problems with the CoMFA method. Both Lennard-Jones and Coulomb potential functions are very steep near the van der Waals surface of the ligand, requiring the use of cut-off values. In addition, entropic contributions to the binding affinities or biological activities are neglected in CoMFA. To overcome these problems, an alternative method CoMSIA[100] was developed by Klebe and co-workers. This approach computes ligands' property fields based on similarity indices of aligned molecules. The similarity is not measured directly by comparing all pairs of molecules. Instead, the similarity is represented by the similarity of each molecule with probe atoms placed at grid lattice points. Gaussian-type functions are selected to calculate the distance dependent similarity indices and to represent different physicochemical properties. Because Gaussian-type functional forms are smooth without singularities, no arbitrary cut-off values are needed. A general CoMSIA procedure is similar to the standard CoMFA approach. A default grid spacing of 1 Å is used (as described in the literature). Similarity indices $A_{F,k}$ between the ligands (j) and a probe atom (at grid point q) are calculated according to the following equation[100]:

$$A_{F,k(j)}^{q} = \sum_{i} w_{probe,k} w_{ik} e^{-\alpha r_{iq}^{2}}$$

where i is the summation index over all atoms of the molecule j; $\omega_{probe,k}$ is the probe atom with a radius of 1 Å, a charge of +1, a hydrophobicity of +1, a hydrogen bond donating of +1, and a hydrogen bond accepting of +1; ω_{ik} is the actual value of the property k of atom i; r_{iq} is the distance between probe atom at the grid point q and atom i of the molecule; and α is the attenuation factor, with a default value of 0.3. The default value is chosen

because a larger α will lead to an increasing attenuation of distance-dependent effects of molecular similarity and global molecular features become less significant. The data table is then constructed from the calculated similarity indices and then the PLS technique is applied to extract a QSAR relationship from the data table. Although CoMSIA also encounters some of the limitations of CoMFA, it has the following obvious advantages:

1) The probes are not limited to steric or electrostatic fields but also include other fields including hydrophobic and hydrogen bonding, which describe more physicochemical properties compared to CoMFA.

2) The entropic effect in the form of hydrophobic interactions can be taken into account at least partially by using hydrophobic probes.[41]

3) CoMSIA contour maps denote the regions occupied by the ligand that favor or disfavor a particular physicochemical property, which provide a more direct visual guide for modification of the ligand.

In our studies, we have applied both CoMFA and CoMSIA to build 3D-QSAR models that illustrate the correlations between specific changes in Rtt109 inhibitor structures and their biological activities. We also expect that information gained from these models will guide the modifications of Rtt109 inhibitors.

5.2 Experimental Section

5.2.1 3D-QSAR Modeling

Initial CoMFA and CoMSIA models were developed using a 115-compound training set (17 actives and 98 inactives from HTS1) comprising all thiourea-based

structures in our in-house compound collection. *In vitro* Rtt109 IC_{50} values for these seventeen actives ranged from 1.73 μM to 11.46 μM according to the HTS1 results. Because molecules with activities spanning about two to three log units are needed to build a statistically sound and reliable model, inactives were included in the training set and were assigned IC_{50} values of 1000 μM.[98] 3D configurations of all 115 structures were generated using geometry optimization by energy minimization in Pipeline Pilot 8.0. Datasets were aligned in SYBYL-x/1.0 using the optimized structure of the most active compound in the training set (**5.1,** GPHR-00029275, Table 5.1) as the template molecule[102] and applying a rigid alignment method based on common substructures. Gasteiger-Hückel[103, 104], Gasteiger-Marsili,[103] MMFF94,[105] and Hückel[104] charges were calculated for all structures. Subsequent removal of outliers decreased the number of active compounds in the training set to 14, and the overall compound count to 93. Following standard CoMFA and CoMSIA procedures, each compound was placed into a three-dimensional lattice with a 2.0 Å grid-point separation. For CoMFA, the steric (van der Waals) and electrostatic (Coulombic) potential energy fields were obtained by summing individual energy interactions between each atom of the compound of interest and an sp^3 carbon probe atom with a charge of +1, located at each grid point. A distance-dependent dielectric function ($\varepsilon = \varepsilon_0 r$, where $\varepsilon_0 = 1$) was applied, and maximum field values were set to 30 kcal/mol for the steric fields and to ±30 kcal/mol for the electrostatic fields. For CoMSIA, similarity indices were computed using an sp^3 carbon with a charge of +1, radius of +1 Å, hydrophobicity of +1, and hydrogen-bond donor and

acceptor properties of +1; the attenuation factor α for the Gaussian-type distance dependence was set to 0.3.

5.2.2 QSAR Regression Analysis

Linear regression equations in QSAR modeling were derived using the partial least squares (PLS) methodology,[106] in which changes in steric and electrostatic fields (CoMFA) and five similarity fields (CoMSIA) were correlated with changes in experimental biological activities (pIC_{50} values) in the training set. The internal predictive ability of the models was first evaluated by "leave-one-out" cross-validation, whereby each compound was removed from the training set, and a new model was derived using the remaining compounds and subsequently used to predict the activity of each extracted compound, yielding the optimum number of principal components (PC) in the linear equation and a cross-validated r^2 (r_{cv}^2) value, which is a statistical representation of internal predictivity. The PLS calculations were then repeated, without cross-validation, using the optimum number of PCs, resulting in final CoMFA and CoMSIA models with corresponding conventional r^2 values, standard errors of estimate, *F* ratios, and other related statistical parameters. Column filtering was applied to any column of computed energies with a variation less than 2.0 kcal/mol, in order to accelerate the regression analysis and to reduce "noise." Each 3D-QSAR model was visualized as a color contour map illustrating regions of descriptor fields that strongly contributed to the overall model.

5.3 Results and Discussion

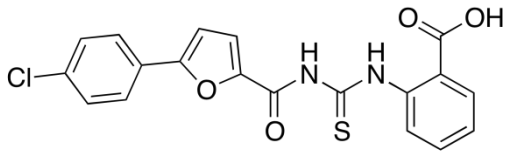
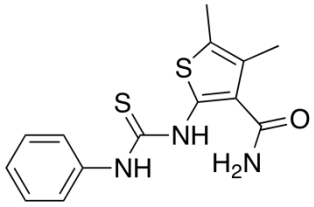
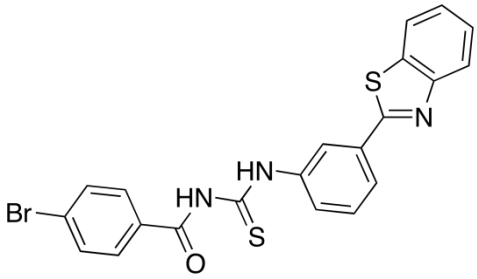
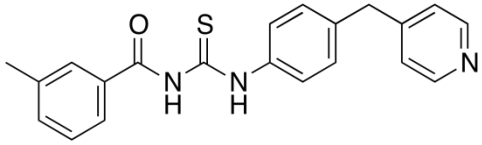
5.3.1. 3D-QSAR Modeling

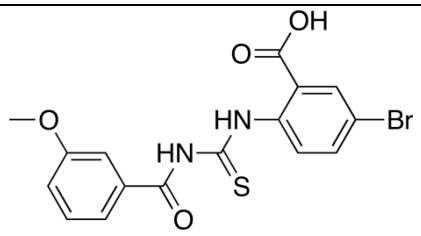
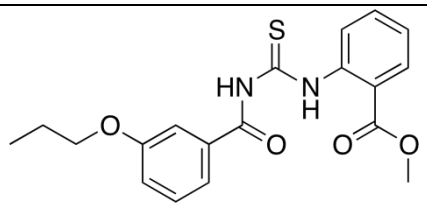
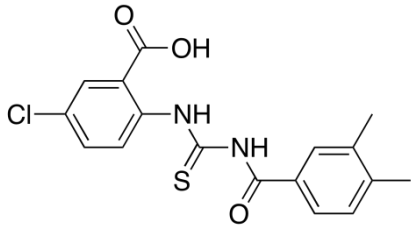
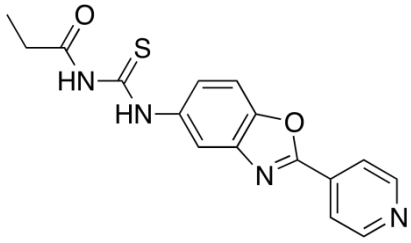
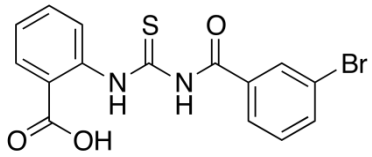
To correlate specific changes in Rtt109 inhibitor structure with biological activities, we generated a series of 3D-QSAR models based on a selection of active and inactive compounds from our HTS1 assay, using CoMFA and CoMSIA. Based on the assumption that enzyme-inhibitor interactions are largely noncovalent, enthalpic, and shape-dependent, these industry-standard methods derive 3D-QSAR models by systematically sampling steric and electrostatic fields surrounding a set of small molecules (CoMFA) or calculating Gaussian-function similarity indices representing steric, electrostatic, hydrophobic, and hydrogen bonding interactions (CoMSIA) and correlating the respective fields with experimental biological activities. Both techniques yield visual representations of areas on the ligands where the respective properties correlate strongly with increases or decreases in biological activity, as well as quantitative models that, if statistically sound, can be used to predict the activities of structurally similar compounds. The predictivity and reliability of the CoMFA and CoMSIA models were assessed by means of four primary parameters: conventional r^2, a measure of how well predicted biological activities for the training set compounds agree with experimentally determined values; cross-validated r^2 (r_{cv}^2, from a "leave-one-out" cross-validation procedure, see 5.2 Experimental Section); *F*-ratio (defined as $r^2/(1-r^2)$ and representing the ratio of properties explained by the QSAR model to those not explained by it[102]); and standard error of estimate, or SEE. After extensive generation and testing of 3D-QSAR models (not reported here) from combinations of various active and

inactive compound subsets, and applying a variety of charge formalisms, a training set was chosen that comprised all compounds in the entire 82,861 in-house compound collection that featured a thiourea core structure (17 actives and 98 inactives), with Gasteiger-Hückel charges.[103, 104] Using this 115-compound thiourea-based set, compound **5.1** (GPHR-00029275, Table 5.1) as the template for molecule alignment, and a QSAR column filtering level of 1.0 kcal/mol, a preliminary CoMFA model was developed that exhibited very good internal predictive ability ($r_{cv}^2 = 0.712$), good self-consistency ($r^2 = 0.938$), and a favorable *F*-ratio (271.225), but also a rather high SEE (0.221). Twenty-two "outlier" compounds (three actives and nineteen inactives) were removed that exhibited residual differences in predicted vs. experimental biological activity greater than or equal to 0.2 log units, resulting in a final training set comprising 14 actives (Table 3) and 79 inactives. (Note that all inactive compounds demonstrated zero activity against Rtt109 and were consequently assigned numerical pIC_{50} values of 3.0 for the purposes of this QSAR analysis – which also enabled the resulting models to distinguish sharply between actives and inactives, given that the training set actives exhibited a rather narrow activity range (1.73 – 11.46 μM) and models based on the training set actives only displayed less satisfactory statistical results). The new CoMFA model derived from this final compound set displayed a slight increase in internal predictivity ($r_{cv}^2 = 0.751$), better self-consistency ($r^2 = 0.985$), a significantly higher *F*-ratio (970.516), and, notably, a decreased standard error of estimate (0.107) (Figure 5.1). The best CoMSIA model obtained using this final training set did not prove as self-consistent as the respective

CoMFA model, but was internally predictive as determined by cross-validation (r_{cv}^2 = 0.651, r^2 = 0.977, SEE = 0.135, and F = 601.876, Figure 5.2).

Table 5.1. Fourteen active, thiourea-based Rtt109 inhibitors from the final 3D-QSAR compound training set, with Rtt109 pIC_{50} and IC_{50} values

Cpd #	Structure	Cpd ID	Rtt109 pIC_{50}	Rtt109 IC_{50} (μM)
5.1		GPHR-00029275	5.76	1.73 ± 0.46
5.2		GPHR-00026554	5.69	2.01 ± 0.35
5.3		GPHR-00006279	5.53	2.96 ± 0.68
5.4		GPHR-00022505	5.49	3.22 ± 0.84

5.5		GPHR-00053162	5.49	3.26 ± 0.60
5.6		GPHR-00023078	5.41	3.89 ± 1.49
5.7		GPHR-00053677	5.36	5.4 ± 1.30
5.8		GPHR-00011814	5.32	5.82 ± 1.38
5.9		GPHR-00029320	5.29	5.04 ± 2.09

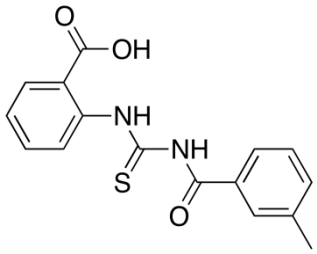
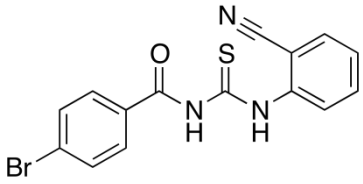
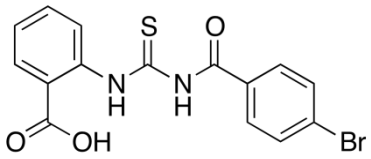
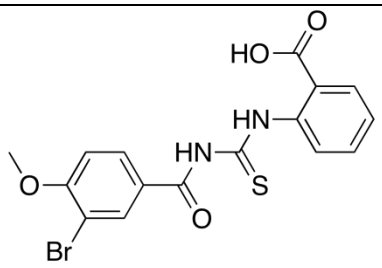
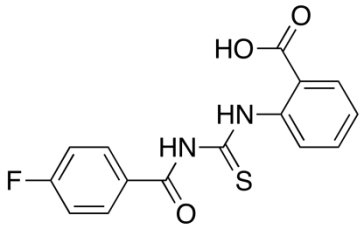
5.10		GPHR-00029429	5.28	5.30 ± 2.14
5.11		GPHR-00058009	5.27	5.32 ± 2.09
5.12		GPHR-00029520	5.19	6.36 ± 2.34
5.13		GPHR-00031245	5.15	7.13 ± 2.30
5.14		GPHR-00029708	4.94	11.46 ± 3.55

Figure 5.1. Plot of the CoMFA-predicted vs experimental biological activities for the final thiourea-based Rtt109 inhibitor training set. All inactive compounds were assigned an experimental pIC_{50} value of 3.0.

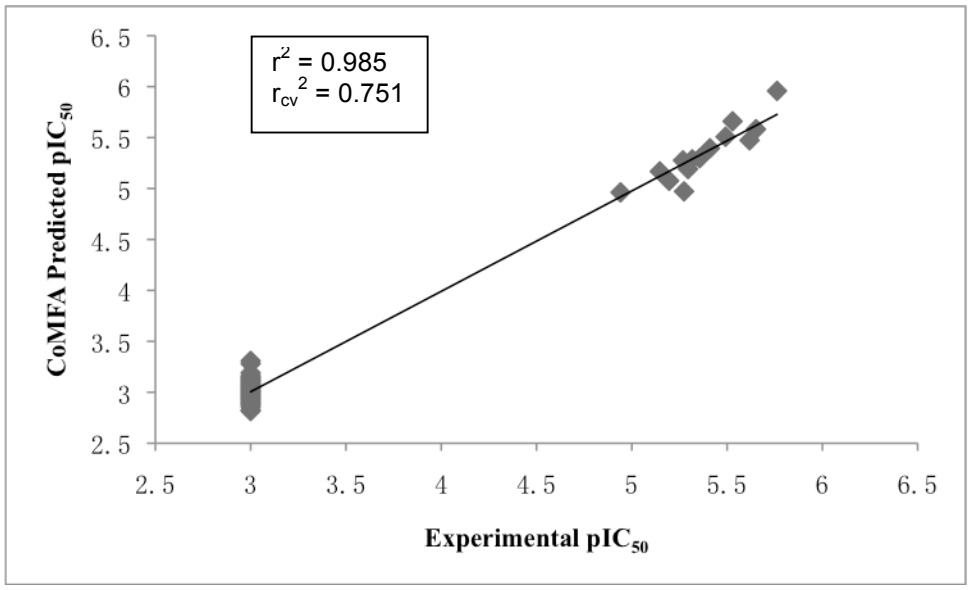
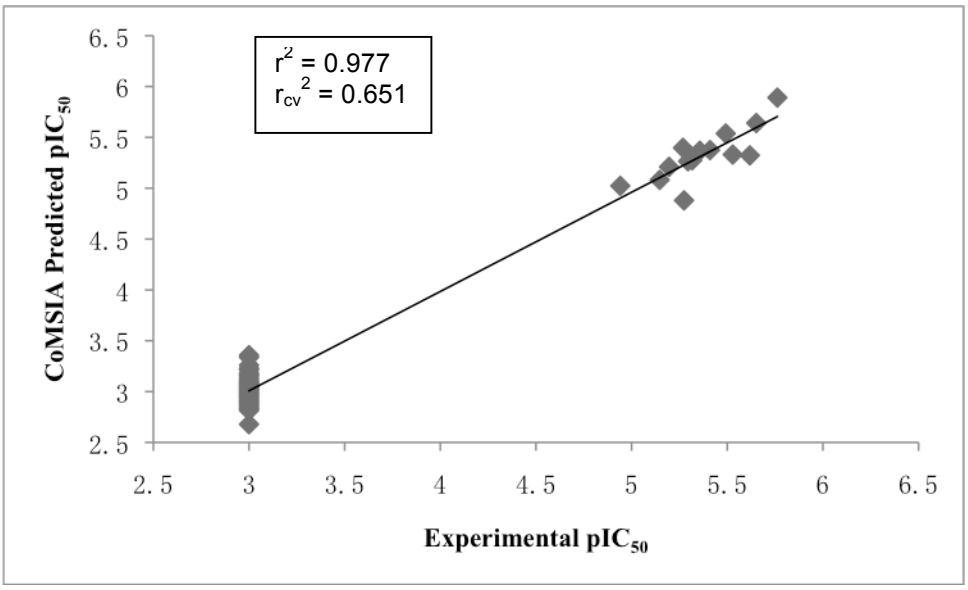


Figure 5.2. Plot of the CoMSIA-predicted vs experimental biological activities for the final thiourea-based Rtt109 inhibitor training set. All inactive compounds were assigned an experimental pIC_{50} value of 3.0.



5.3.2 Model Interpretation

The color contour map of the most predictive CoMFA model is shown in Figure 5.3, together with the docked configuration of compound **5.1,** superimposed on the MOLCAD[107, 108] fast Connolly electron-density surface of the Rtt109 Lys-AcCoA binding tunnel, with electrostatic potential mapping. All training set active compounds were docked using the same method and displayed similar docking poses compared to the docking pose of compound 5.1. Green polyhedra indicate regions on the inhibitor where steric bulk is predicted to favor biological activity against Rtt109, whereas yellow polyhedra denote areas where ligand bulk is likely to decrease activity. Similarly, blue and red areas represent, respectively, inhibitor areas where positive and negative electrostatic potential correlate with increased activity. A broad green region located adjacent to the *p*-chlorophenyl moiety in **5.1** suggests a tolerance for steric bulk at the AcCoA binding region of the tunnel. Yellow polyhedra located nearby highlight steric constraints that come into play as the tunnel narrows. Favorable steric potential as denoted by a green area near the *o*-carboxylphenyl grouping appears to conflict with corresponding structural features in the receptor. However, it more accurately reflects the variety of bound configurations possible at the entrance to the Lys tunnel rather than a need for bulky groups to be added to the ligand in this area. A strong preference for electronegative groups near the carboxylic acid moiety on **5.1** (red polyhedra) is consistent with the observation from our *in vitro* HTS1 data that negatively charged functionalities on one or more of the ligand aromatic moieties generally correlate with an

increase in Rtt109 inhibition. Two smaller blue regions near the hydrogen atoms of the *o*-carboxyphenyl moiety point to a preference for electropositive groups in these areas.

Figure 5.3. CoMFA contour map (SYBYL-x/1.0, Tripos, Inc.) for the final thiourea-based Rtt109 inhibitor training set, shown with compound **5.1** (Rtt109 pIC_{50} = 5.762, IC_{50} = 1.73 μM) and the MOLCAD electron-density surface of the Rtt109 Lys-Ac-CoA binding tunnel, with electrostatic potential mapping on the receptor (red = positive; violet = negative). Colored polyhedra represent areas on or near the ligand where properties correlate strongly with biological activities, where red = negative electrostatic potential; blue = positive electrostatic potential; yellow = negative steric potential; green = positive steric potential.

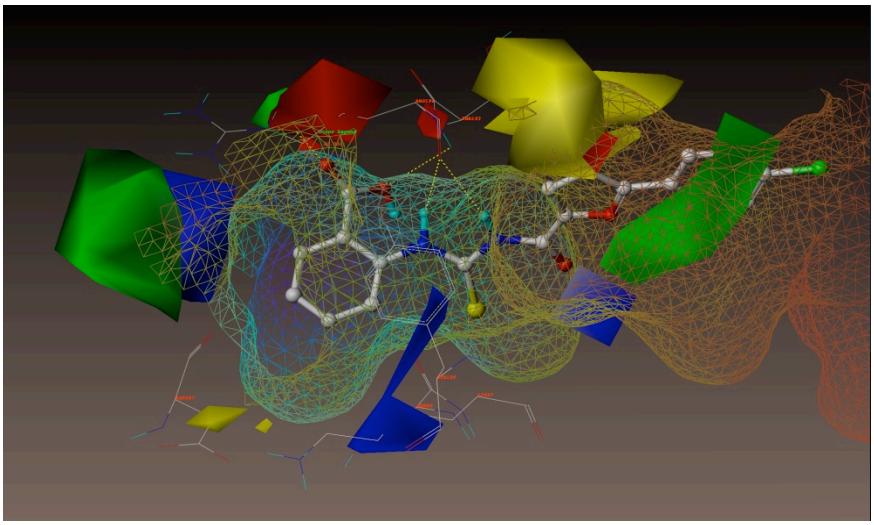


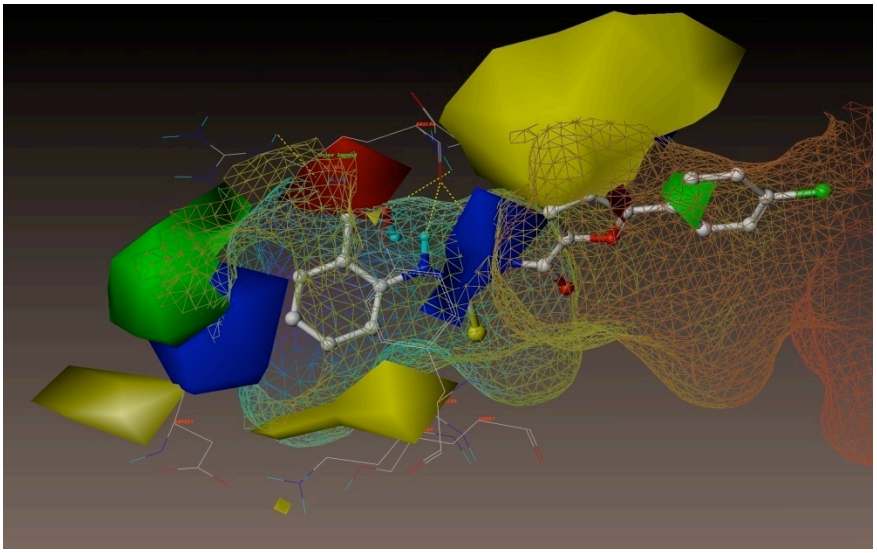
Figure 5.4 illustrates the corresponding CoMSIA model, shown with compound **5.1** and the MOLCAD[107, 108] Rtt109 binding tunnel electron density surface with electrostatic (Figure 5.4a), hydrogen-bonding (Figure 5.4b) and lipophilic (Figure 5.4c) potential mapping. The CoMSIA steric and electrostatic fields agree well with those obtained by CoMFA, and the additional hydrogen-bond donor, hydrogen-bond acceptor and hydrophobic contour maps provide supplemental information regarding small-molecule binding requirements in the Rtt109 Lys tunnel. In Figure 5.4b, a cyan area located proximal to thiourea amino groupings on the ligand and near the backbone carbonyl of Arg194 on the receptor indicates that a hydrogen-bond donor in this position will correlate with an increase in biological activity. A corresponding magenta polyhedron (H-bond acceptors favored) in the vicinity of the ligand carboxylic acid points to a key hydrogen bonding interaction between that carbonyl and one of the Arg194 side chain amino moieties, as well as between the Thr193 backbone amino grouping and the carboxyl OH of the inhibitor. Interestingly, the CoMSIA fields represent proportionally more disfavored interactions in this case than the CoMFA maps, perhaps due to five fields (rather than two) being calculated for the relatively large number of inactive compounds in the training set. For example, in Figure 5.4b, the large purple area on the lower left-hand side overlaps with a yellow polygon, underscoring a correlation between hydrogen-bond donors, steric bulk, and decreased biological activity in this region. In addition, a low tolerance for hydrogen-bond acceptors on the ligand near the Rtt109 Pro297 side chain is denoted by the orange polygon on the upper left-hand side of the diagram, while the nearby (and slightly overlapping) purple area indicates that

hydrogen-bond donors are also disfavored at this region – both of which, taken together, correspond with structural biology data that hydrogen bonding of any kind is not likely to take place in this part of the tunnel. Another orange region near the lower left-hand side of compound **5.1**, flanked by the large purple area, indicates that hydrogen-bond acceptors are disfavored on the ligand near the aliphatic part of Lys87, which is further consistent with experimental structural data as no hydrogen-bond donors within an acceptable distance are present on the receptor.

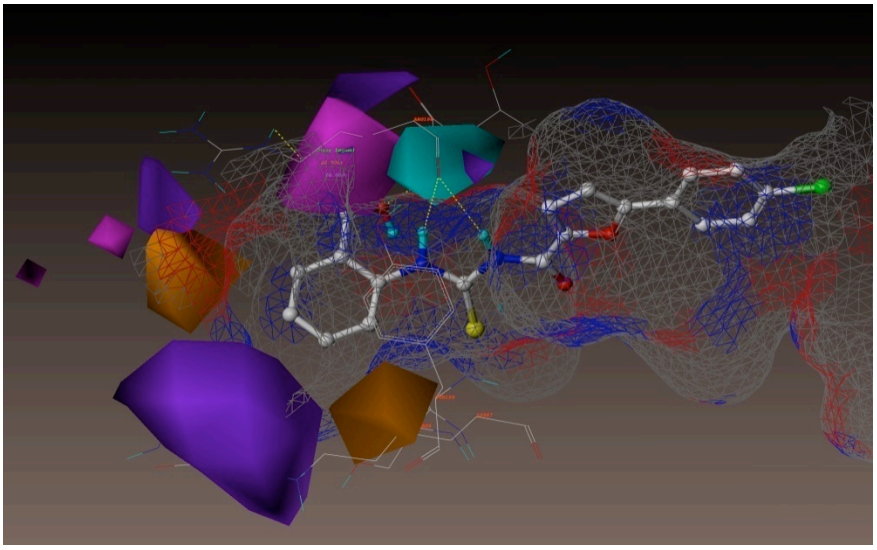
The hydrophobic contour map of the best CoMSIA model (Figure 5.4c) correlates very well with hydrophobic features in the putative Rtt109 small-molecule binding site. The hydrophobic part of the Lys-AcCoA tunnel formed by residues Lys210, Ile212, and Trp221 is correctly reflected by the two brown polyhedra in the contour map. A gray area near the upper right-hand of the molecule suggests that hydrophobicity may be disfavored at that location, but viewing the hydrophobic and steric similarity indices simultaneously reveals an overlap between this gray polygon and an area where steric bulk is strongly disfavored, meaning that no substituents of any kind are allowed in that region, hydrophobic or otherwise. However, the gray polyhedron on the ligand near the solvent-exposed entrance to the tunnel (lower left side of the illustration) correlates with the more hydrophilic environment in this area. Good agreement was observed between all features in the best CoMFA and CoMSIA models and steric, electrostatic, hydrophobic and hydrogen-bonding requirements in the Rtt109 Lys tunnel as determined by experiment.

Figure 5.4. CoMSIA contour maps (SYBYL-x/1.0, Tripos, Inc.) for the final thiourea-based Rtt109 inhibitor training set, shown with predicted bound configuration of compound **5.1** (Rtt109 pIC_{50} = 5.762, IC_{50} = 1.73 μM) and the MOLCAD electron-density surface of the Rtt109 Lys-Ac-CoA binding tunnel, with (a) electrostatic potential mapping on the receptor and steric/electrostatic fields on the ligand; (b) hydrogen bonding potential mapping on the receptor and hydrogen-bond donor and acceptor fields on the ligand; and (c) lipophilic potential mapping on the receptor and hydrophobic fields on the ligand. For receptor surface mapping: (a) red = positive electrostatic potential; blue = negative electrostatic potential; (b) red = H-bond donor, low electronegativity; blue = H-bond receptor, high electronegativity; gray = no H-bonding; (c) brown = highest lipophilicity; blue = highest hydrophilicity. For colored polyhedra on the ligand: (a) red = negative electrostatic potential; blue = positive electrostatic potential; yellow = negative steric potential; green = positive steric potential; (b) cyan = H-bond donors favored; purple = H-bond donorsdisfavored; magenta = H-bond acceptors favored; red = H-bond acceptors disfavored; and (c) brown = high hydrophobicity; gray = high hydrophilicity (or hydrophobicity disfavored).

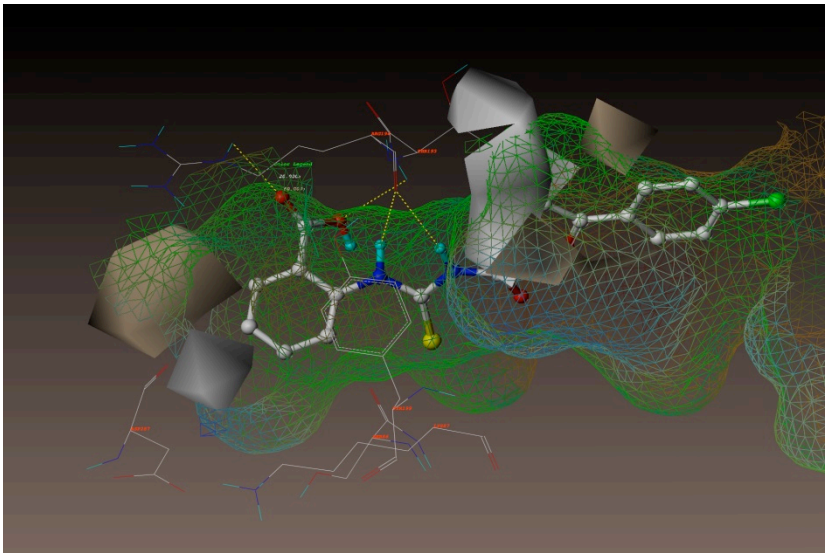
5.4 (a)



5.4 (b)



5.4 (c)



5.4 Concluding Remarks

With the available activity data at the time when we built the models, predictive and self-consistent CoMFA and CoMSIA models were derived for a series of thiourea-based compounds; these 3D-QSAR hypotheses were expected to prove useful for the design and refinement of structurally related compounds and are expected to predict biological activities of similar structures with sufficient accuracy. Although in the follow-up interference assays and orthogonal slot blot assays, the majority of the thiourea-based compounds used to build the models may have been identified as false-positives, our 3D-QSAR analysis yielded color contour maps that match structural features in the experimental Rtt109 X-ray structure quite well, most notably highlighting steric constraints in the Lys-AcCoA binding tunnel and hydrophobic requirements for Rtt109 inhibition, and pointing to key hydrogen-bonding interactions with Thr193 and Arg194 that are expected to increase activity.

CHAPTER 6:

PHARMACOPHORE MODELING STUDIES ON RTT109 INHIBITORS

6.1 Introduction

Pharmacophore is an abstract representation of the key features common to a set of active compounds that interact with the complementary sites on the corresponding receptor.[109] Pharmacophore mapping has proven useful for identifying chemical functionalities responsible for biological activity and for pinpointing new drug and probe scaffolds via chemical database searching.[109] Successful stories can be found in many medicinal chemistry publications, including the discoveries of the Type 1 11-β-Hydroxysteroid dehydrogenase (11-β-HSD1) inhibitors,[110] Hepatitis C virus NS5B protein inhibitors,[111] hormone sensitive lipase inhibitors,[112] anthrax lethal factor inhibitors,[113] and compounds with strong affinity for various G-coupled protein receptors.[114] Pharmacophores can also be applied to model the disruption of protein-protein interactions. For example, without knowing the exact structural information related to the protein-protein interactions, pharmacophore models generated based on known disruptors of the c-Myc-Max dimer were able to identify nine novel disruptors in *in vitro* and in cellular assays.[115]

6.1.1 GALAHAD Methodology

Various pharmacophore elucidation programs are available and widely used, including CATALYST,[116] GALAHAD,[117] the pharmacophore module of MOE[118], and PHASE.[119] All were designed to address three core problems in the pharmacophore

elucidation: 1) representing the pharmacophoric features, 2) searching for potential alignments among the input active compounds, and 3) scoring and ranking these alignments.

GALAHAD is one of the pharmacophore elucidation algorithms we used in our studies. It is short for a Genetic Algorithm with Linear Assignment for Hyper-molecular Alignment of Datasets.[117, 120] In GALAHAD, the default pharmacophoric feature types are hydrogen bond donor and acceptor atoms (D and A), positive nitrogen (P), negative and hydrophobic centers (N and H), and steric features (S). GALAHAD divides the pharmacophore elucidation into two stages. It first utilizes a genetic algorithm (GA) to explore the ligand torsional space and to find a set of ligand conformations that have low steric energy, large common volume/shape, and high similarity in pharmacophoric features. GALAHAD then uses the extended LAMDA[121] (Linear Assignment for Molecular Dataset Alignment) algorithm to perform a final alignment of ligands, of which conformations are held rigid from the first step. In the first stage, a multi-objective fitness Pareto function in the GA process is used to produce a population of possible models, each representing a different trade-off between the competing criteria (strain energy, volume overlap, and feature matching).[122] The Pareto scoring function is a significant improvement over functions implemented in other earlier pharmacophore mapping algorithms. During the second stage, the LAMDA algorithm is extended beyond just a pair-wise atom-based alignment method. The extended LAMDA aligns datasets of two or more molecules based on shared features and the generation of a hyper-molecule and it does not require a template molecule for alignment. A hyper-molecule of a dataset

is produced by overlaying pairs of molecules or hyper-molecules and merging features of the same type that lie closely (the default threshold distance being 0.6 Å) while preserving the connectivity and geometry of each molecule in the dataset. This hyper-molecule identifies the most important common substructures and conformations between a set of molecules and provides alignment rules for molecules not previously included in its generation.

In our studies, we used GALAHAD as the primary pharmacophore mapping tool to identify key features of the active compounds that contribute to their activities against Rtt109.

6.2 Experimental Section

6.2.1 Pharmacophore Modeling

In HTS1 (See Chapter 2), 213 compounds (dataset **ITB**) were confirmed via dose-response to be active against Rtt109, with activity ranging from 0.49 – 17.5 μM, while the 137 compounds displayed questionable dose-response curves and were classified as inactive. A structurally diverse subset of two hundred small molecules was selected (by Tanimoto coefficient[123]) from among the 82,648 inactives (82,511 compounds from the primary screen which displayed <50% inhibition, and 137 designated inactives from the dose-response screen) and was assembled into validation control database **ITC.**

The 213 active compounds from our experimental HTS1 were clustered into eleven subgroups (datasets **ITB1-ITB11**) according to Tanimoto similarity with an average number of 20 compounds per cluster using SciTegic Pipeline Pilot 8.0 (Accelrys,

Inc.). Clustering was done by maximizing Tanimoto distance between cluster centers while minimizing that distance within each cluster. As described above, the two hundred compounds comprising inactive control dataset **ITC** were chosen from HTS-identified inactives using Tanimoto-based diverse subset selection in Pipeline Pilot. Three-dimensional conformations of all dataset structures were generated via geometry optimization by energy minimization in Pipeline Pilot. All pharmacophore models were generated, analyzed, and optimized using genetic algorithms and Pareto scoring as implemented in the GALAHAD module (SYBYL-x/1.0). Models were obtained by iterating for a maximum of 40 generations with a population size of 35. A 40% mutation rate per torsional angle gene and a 100% crossover rate were applied in each GALAHAD run. Two key aromatic features were required to be present in all models in order to maximize enrichment index (EI) values; among remaining features, partial matching was enabled so that identified compounds must match at least the total number minus one (N-1) of all features.

6.3 Results and Discussion

6.3.1 Pharmacophore Modeling

The 213 active compounds comprising dataset **ITB** were divided into 11 subsets (datasets **ITB1 – ITB11**) based on structural similarity as measured by Tanimoto coefficient.[35] Seven of these subgroups comprised more than five structures; therefore, the five most active compounds, based on *in vitro* HTS1 Rtt109 IC_{50} values, from each of these groups were selected to create training sets to develop pharmacophore hypotheses

for each group. Additionally, the five most active compounds overall, i.e., across all eleven clusters in dataset **ITB**, were assembled into a general training set. Multiple pharmacophore hypotheses were generated for the seven clusters and for the general training set using genetic algorithms (GAs) and a scoring function based on Pareto multi-objective optimization, as coded in the GALAHAD software module.[117, 120, 121] For each hypothesis, a number of key criteria were obtained: specificity (a logarithmic indicator of the expected discrimination for each query, based on the number of features in the model, their partial match constraints, and distribution of the features in space);[102] the number of training set compounds matching the pharmacophore; steric overlap; pharmacophoric concordance; and strain energy.[120, 124] Models with high specificity values (and broader spatial distribution of features) are preferred, as nonspecific or ambiguous hypotheses generally cannot serve as effective database queries. The accuracy of each model was assessed by a multi-objective triage approach, in which hypotheses were first assigned a Pareto score, which is the number of other models obtained for the same training set that are judged superior to it in terms of total strain energy, steric overlap, pharmacophoric concordance, and tuplet agreement. If Pareto scores were identical, models were then ranked by Borda tally across pharmacophoric and steric multiplet consensus terms.[102] In exceptional cases where top-ranked models (according to Borda sum) may have been questionable because they only matched two or three training set compounds, models were instead prioritized using specificity and number of compounds matched as key criteria. Any remaining "ties" were then broken in favor of hypotheses with lower overall energy.

6.3.2 Model Validation

All eight final hypotheses – the general model, plus seven models from each compound subset – were validated by means of virtual database screening, in which the hypotheses were represented as UNITY queries (SYBYL 8.0, Tripos, Inc.) and subsequently used to search active compound database **ITB** and inactive control database **ITC.** In the database screen, partial match criteria were established (see 6.2 Experimental Section above), as requiring all features in a given model to be matched in order for a given compound to be considered a "hit" may result in false negatives, especially in the current case where the active Rtt109 inhibitors are structurally diverse and may interact with various sub-regions in the target binding area. The predictive ability of each hypothesis was assessed by the number of actives and inactives returned by that model, and by enrichment index **EI**, which represents the ratio of correctly identified actives to incorrectly identified inactives:

$$EI = \frac{I_{active}}{D_{active}} \bullet \frac{D_{inactive}}{I_{inactive}} \quad (1)$$

where I_{active} is the number of active compounds returned by the pharmacophore model query; $I_{inactive}$ is the nonzero number of inactive compounds returned by the query, i.e., incorrect matches; D_{active} is the number of structures in the active compound database **ITB** (213); and $D_{inactive}$ is the number of structures in the inactive control database **ITC** (200).[125] Results from this screen are shown in Table 6.1: the four hypotheses **H1-H4** (derived from the general training set and datasets **ITB1, ITB5,** and **ITB11,** respectively) returned few or no inactive compounds. (Four models from the remaining structural subgroups **ITB3, ITB4, ITB6** and **ITB7** erroneously matched a large number of

inactives, i.e., were nonspecific, and were therefore excluded from further consideration.) The general model **H1** retrieved the highest number of actives (107, 50.23% of **ITB** compounds) from among the four feasible hypotheses, but also returned 12 false positives (6% of **ITC** structures), resulting in a relatively low enrichment index of 8.37. Model **H2** yielded a much higher enrichment index (30.99) but was able to identify only thirty-three active compounds. Similarly, hypothesis **H3** from dataset **ITB5** demonstrated the highest enrichment index (36.62), identifying only one inactive compound out of 200, but returned only thirty-nine actives. Note that an enrichment index could not be calculated for model **H4**, as it returned no inactive structures; however, this hypothesis pinpointed even fewer actives: only twenty out of 213. Model **H1** was therefore chosen as the preferred pharmacophore hypothesis.

Table 6.1. Pharmacophore model validation results, obtained by screening active compound database **ITB** and inactive compound database **ITC** using four hypotheses **H1-H4**

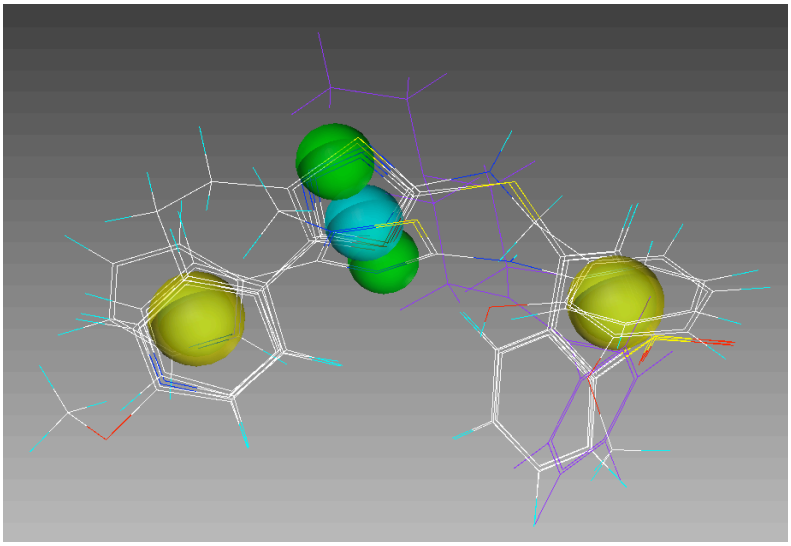
Model	Actives Matched (Number, Percent of Total)	Inactives Matched (Number, Percent of Total)	EI[a]
H1	107/213, 50.23%	12/200, 6.00%	8.37
H2	33/213, 15.49%	1/200, 0.50%	30.99
H3	39/213, 18.31%	1/200, 0.50%	36.62
H4	20/213, 9.39%	0, 0.00%	N/A

[a]Enrichment index

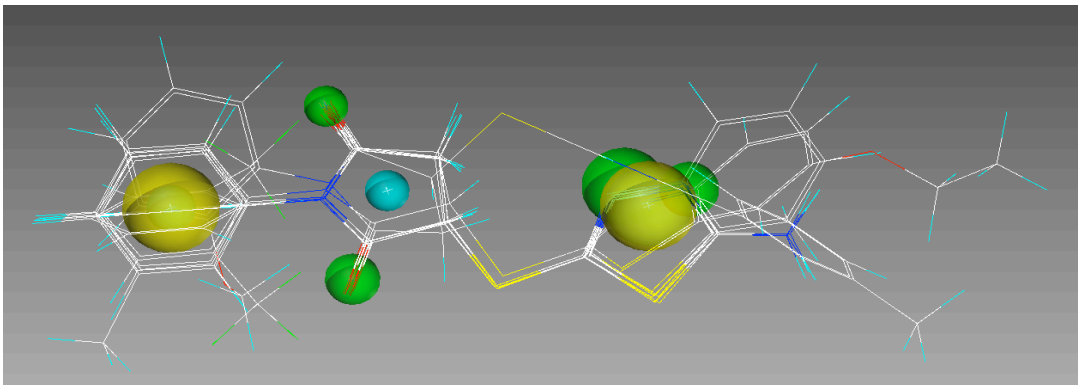
Graphical renderings of models **H1-H4** are shown in Figure 6.1. The general model **H1** (Figure 6.1a), derived from the five most active Rtt109 inhibitors we tested, comprises five features: two hydrogen bond acceptors, two aromatic/hydrophobic features and one hydrophobic feature. The other three models (**H2-H4,** Figure 6.1b-6.1d), all incorporate the two hydrophobic/aromatic features seen in model **H1,** but vary significantly in terms of hydrogen-bond donor and acceptor positioning, which is not surprising given the distinct structural dissimilarity across the compound training subsets. Note, however, that a centrally positioned hydrophobic feature is present in all models except **H3** (although the distance between this feature and the two hydrophobic/aromatic features varies somewhat, especially in the case of model **H4**). These shared pharmacophoric features suggest that the compounds in all four of these subgroups, although structurally diverse, may exhibit similar Rtt109 binding modes.

Figure 6.1. Pharmacophore hypotheses generated from four active Rtt109 inhibitor subsets: (a) **H1,** from the general training set; (b) **H2,** from subset **ITB1**; (c) **H3,** from subset **ITB5**; and (d) **H4,** from subset **ITB11**, shown with representative structures used to derive the models. (SYBYL-x/1.0, Tripos, Inc.) Features are colored as follows: cyan = hydrophobic; green = hydrogen-bond acceptor; magenta = hydrogen-bond donor; yellow = aromatic).

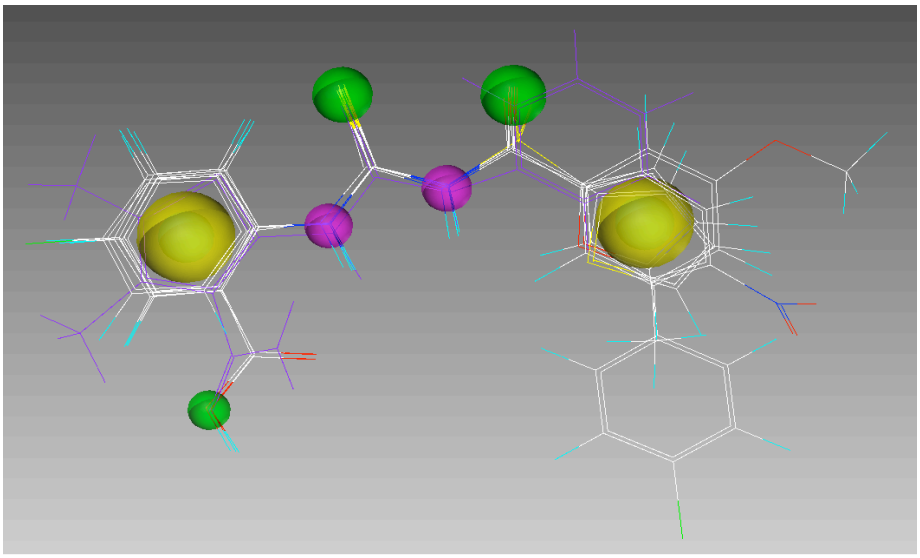
6.1 (a)



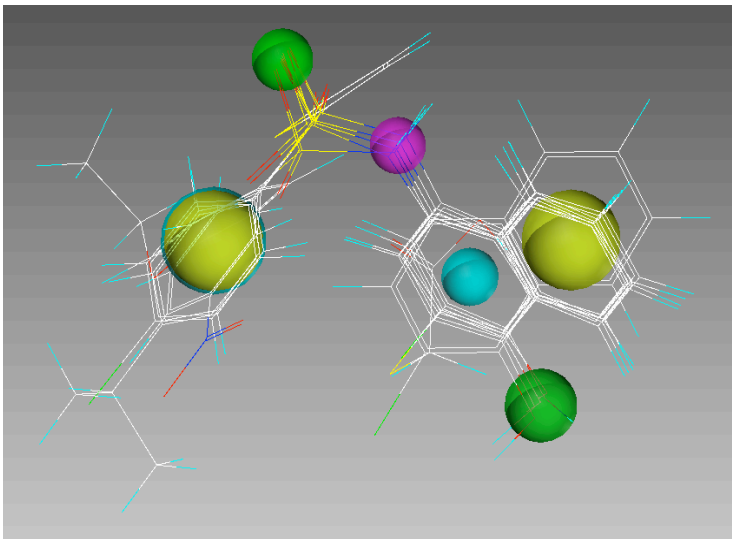
6.1 (b)



6.1 (c)



6.1 (d)

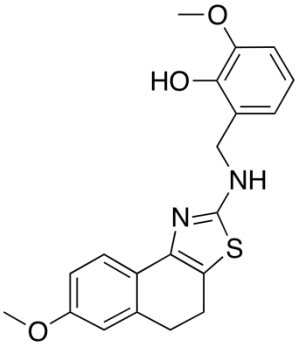
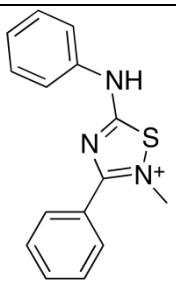
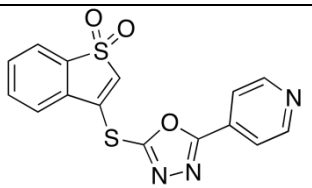
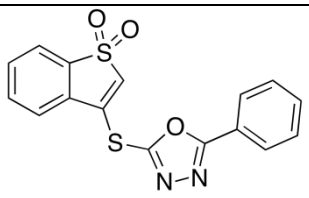


6.3.3 Model Complementarity with the Rtt109 Active Site and 3D-QSAR Models

To assess the compatibility of model **H1** with experimentally determined structural features in the lysine-AcCoA tunnel, the pharmacophoric features of this model were superimposed upon this area in the Rtt109 X-ray structure (2ZFN.pdb.[31]) The most active compound (**6.1,** GPHR-00049940, see Table 6.2) identified by our HTS1 assay was docked into the Lys-AcCoA binding tunnel (Surflex-Dock, SYBYL 8.0, Tripos, Inc.), and its predicted bound conformation is illustrated in Figure 6.2 together with model **H1** and nearby residues in the Rtt109 crystal structure. All three demonstrate good agreement, further supporting the hypothesis that the inhibitors pinpointed in our experimental screen bind to Rtt109 in the lysine tunnel area. The hydrophobic/aromatic feature HY2_ARO6 in model **H1** represents moieties that can engage in hydrophobic interactions with nearby aliphatic portions of the Rtt109 Lys87 and Arg194 side chains, as well as with Tyr199 and Phe84, and this feature may also signify a π-π interaction with Tyr199. The hydrogen-bonding features AA1 and AA4 are found in the vicinity of Trp222 and Ala88, respectively, indicating that inhibitor activity may be enhanced by hydrogen-bonding groups in those locations that can interact with the amino moieties of Trp222 and Ala88. Furthermore, the hydrophobic/aromatic feature HY_5_ARO7 points to one or more inhibitor functionalities that can accommodate the hydrophobic environment formed by Lys210, Ile212, Leu218, and Trp221. Notably, key residues such as Arg194, Lys210, Ile212, and Trp221, pinpointed by the pharmacophore models were also identified by 3D-QSAR models as the important residues for inhibitor binding. The

consistency between the 3D-QSAR models and pharmacophore maps serves as supporting evidence that the compounds may bind to the Lys-AcCoA tunnel of Rtt109.

Table 6.2. General training set comprising the five most active compounds overall from dataset **ITB,** used to generate pharmacophore model **H1**, with Rtt109 activity values

Compound #	Structure	Compound ID	Rtt109 IC_{50} (μM)
6.1		GPHR-00049940	0.49 ± 0.02
6.2		GPHR-00032548	0.71 ± 0.05
6.3		GPHR-00052071	0.90 ± 0.21
6.4		GPHR-00054718	0.91 ± 0.22

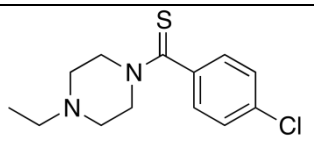
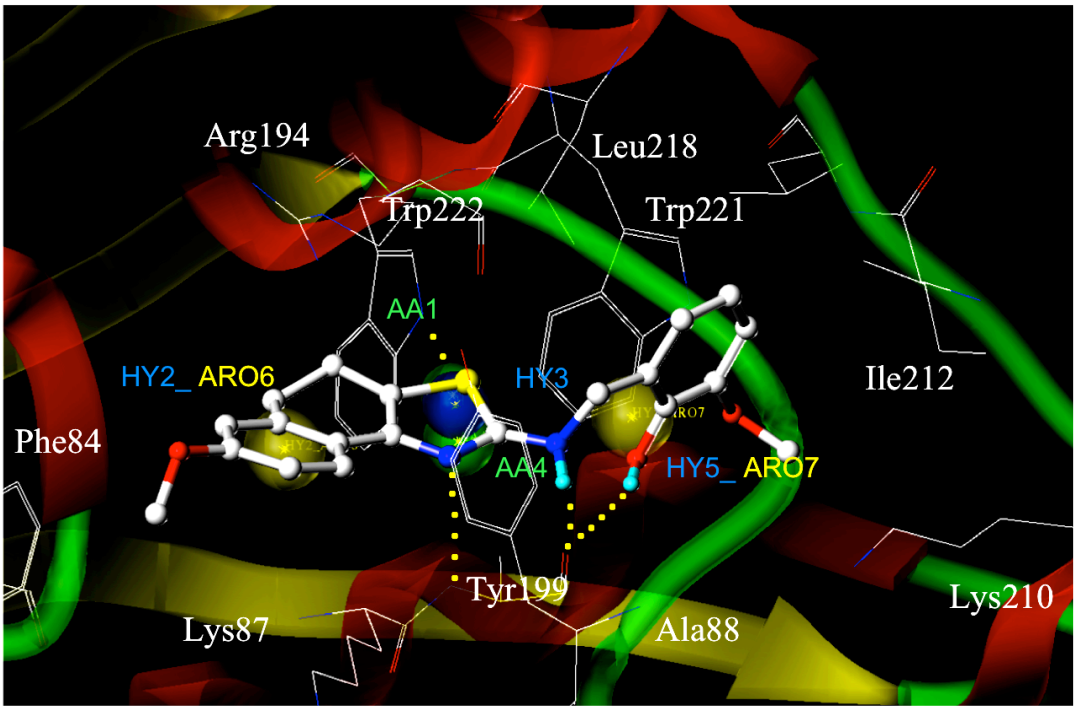
6.5		GPHR-00050689	0.97 ± 0.29

Figure 6.2. Final Rtt109 inhibitor pharmacophore model **H1** based on the five most potent compounds overall as identified by experimental HTS, shown with the predicted bound conformation of most active compound **6.1**(GPHR-00049940) and nearby residues in the Rtt109 Lys-Ac-CoA binding region (2ZFN.pdb[31]). (SYBYL-x/1.0, Tripos, Inc.) Features are colored as follows: cyan = hydrophobic; green = hydrogen-bondacceptor; magenta = hydrogen-bond donor; yellow = aromatic).



6.4 Concluding Remarks

Novel, genetic algorithm-based pharmacophore hypotheses were developed and validated using various subsets of active compounds based on HTS1, which elucidate key hydrogen bonding and hydrophobic/aromatic requirements for successful Rtt109 inhibition, and which can also be used to search *in silico* compound databases to identify additional novel scaffold structures. With further studies of the active compounds, including the follow-up interference assays and orthogonal slot blot assays, the majority of the active compounds used to build the models may have substequently been identified as false-positives. However, the pharmacophore hypotheses point to critical receptor residues demonstrated by experiments, including Tyr199 and Trp222 that are predicted to participate in small-molecule binding, which may provide insight into potential mechanisms of Rtt109 inhibition. Most notably, features in the best pharmacophore hypotheses and 3D-QSAR models derived from our HTS1 actives exhibit high complementarity, and closely match steric, electrostatic, hydrogen-bonding and hydrophobic features in the Rtt109 X-ray structure (2ZFN.pdb[31]), supporting the hypothesis that these compounds bind at the lysine entrance tunnel. It will nevertheless be important to experimentally confirm our predicted Rtt109 inhibitor binding modes via compound cocrystallization.

CHAPTER 7:

CLASSIFICATION OF HIGHLY UNBALANCED RTT109 ACTIVITY DATA USING A COST SENSITIVE SUPPORT VECTOR MACHINE TECHNIQUE

7.1 Introduction

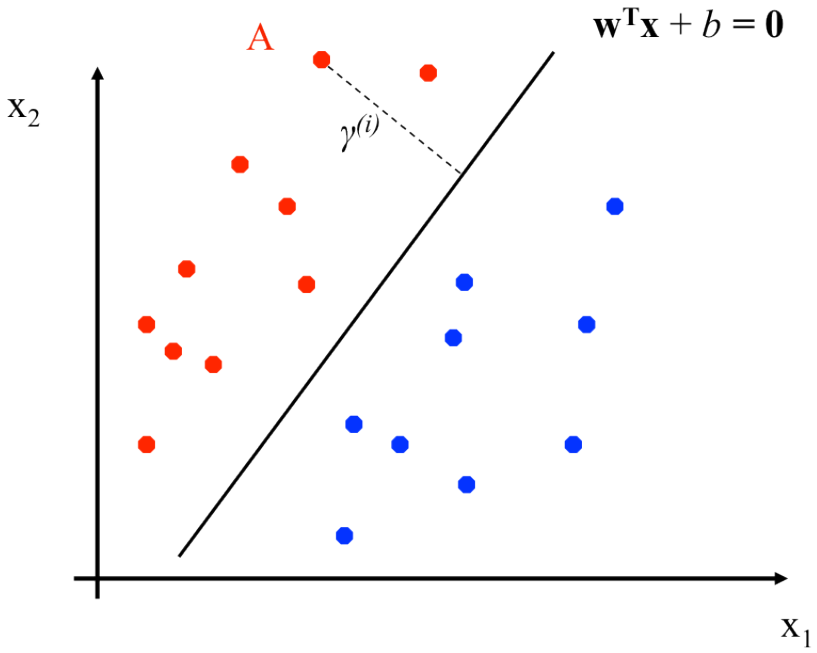
The Support Vector Machine (SVM) is a popular and effective classification algorithm. Intuitively, the general idea behind SVM is to present the examples as points in the feature space and separate them with a large gap (margin).[126] In the cheminformatics field, the features that represent compounds are usually their structural and physicochemical properties, including molecular weight, logP, area of van der Waals surface, polar surface area, connectivity and shape indices, and so on.

7.1.1 Geometric Margins and the General SVM Algorithm

In the following figure Figure 7.1, red dots represent positive training examples ($y^{(i)} = 1$), e.g., the actives, and blue dots denote negative training examples ($y^{(i)} = -1$), e.g., the inactives. The features X_1 and X_2 can be the compounds' molecular weight and logP. A decision boundary ($\omega^T x+b = 0$) is also shown. The point at A represents the input $x^{(i)}$ of positive training examples. Its distance to the decision boundary is denoted as $\gamma^{(i)}$, which is the geometric margin of (ω, b) with respect to a training example ($x^{(i)}$, $y^{(i)}$). The distance can be expressed as the following:

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{\omega}{\|\omega\|}\right)^T x^{(i)} + \frac{b}{\|\omega\|}\right)$$

Figure 7.1. Illustration of a decision boundary (separating hyperplane)



For example, the geometric margin of (ω, b) with respect to the training set S = $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ is then defined as the smallest geometric margin on the individual training examples:

$$\gamma = \min_{i=1, \ldots, m} \gamma^{(i)}$$

The goal of the SVM algorithm is then to find a decision boundary that maximizes the geometric margin, which reflects a powerful classifier that can be used to make confident and correct predictions on the training examples.[126, 127] This can be formalized as an optimization problem as follows:

$$\max_{\gamma, \omega, b} \gamma$$
$$\text{s.t. } y^{(i)} \left(\omega^{T} x^{(i)} + b\right) \geq \gamma, \quad i = 1, \ldots, m$$
$$\|\omega\| = 1.$$

i.e., maximize the geometric margin γ, subject to the condition that each training example has geometric margins of at least γ. However, the "||ω|| = 1" constraint is non-convex and cannot be solved with the standard optimization method. Thus the problem needs to be transformed into the following one:

$$\max_{\gamma, \omega, b} \frac{\hat{\gamma}}{\|\omega\|}$$
$$\text{s.t. } y^{(i)} \left(\omega^{T} x^{(i)} + b\right) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

By introducing a scaling constraint that $\hat{\gamma} = 1$, then the optimization can be further transformed:

$$\min_{\gamma, \omega, b} \frac{1}{2} \|\omega\|^2$$
$$\text{s.t. } y^{(i)} \left(\omega^{T} x^{(i)} + b\right) \geq 1, \quad i = 1, \ldots, m$$

Sometimes, the data may not be separable. In some other cases where outliers exist, finding a separating hyperplane may not offer us the best solution to our problem. In order to make the algorithm work for non-separable data and less sensitive to outliers, the optimization problem can be formulated with regularization terms:

$$\min_{\gamma,\,\omega,\,\mathrm{b}} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{s.t. } y^{(i)}\left(\omega^{\mathrm{T}}x^{(i)} + b\right) \geq 1 - \xi_i,\ i = 1,\ ...,\ m$$
$$\xi_i \geq 0,\quad \mathrm{i} = 1,\ ...,\ m.$$

A cost Cξi will be paid for a mislabeled example. The parameter C determines the relative weighting between both the goals of keeping $\|\omega\|^2$ small and of making sure that most examples can be predicted correctly (with geometric margin at least 1), namely a trade-off between a large margin and a small classification error penalty.[127] By using the Lagrangian multipliers, we obtain the dual form of the optimization problem:

$$\max_{\alpha}\ W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\left\langle x^{(i)}, x^{(j)}\right\rangle$$
$$\text{s.t. } 0 \leq \alpha_i \leq C,\quad i = 1,\ ...,\ m$$
$$\sum_{i=1}^{m}\alpha_{\mathrm{i}}y^{(i)} = 0.$$

This dual problem can then be solved using computational tools.

7.1.2 Kernels

The input attributes of a problem can be mapped to some new sets of features that will be used in a classification algorithm. With a feature mapping expressed as ϕ, the corresponding kernel is defined as follows:

$$K(x,z) = \phi(x)^T\phi(z)$$

All the inner products <x, z> in the previously defined optimization problems can now be replaced with K(x,z). The advantage of this replacement is that K(x,z) may be inexpensive to calculate, even though the calculation of ϕ(x) may be very costly due to the fact it can be a high dimensional vector.[127] Thus by using an efficient way to calculate K(x,z), SVM algorithms can classify examples in high dimensional space given the feature mapping ϕ, without explicitly representing vectors ϕ(x). This allows SVM algorithms to separate non-linearly separable data in a higher dimensional space.

In the computer aided drug design field, the SVM algorithm can be used to generate ligand-based binary classification models. Unlike 3D QSAR models that usually depend on structurally related compounds to produce reliable results and unlike pharmacophore models that often rely on the information of the most active compounds, SVM models take much more experimental information into account by utilizing structural and activity information of all actives and inactives. Thus, we have built SVM classification models that complement our 3D-QSAR and pharmacophore models and that serve as an additional tool for identifying novel Rtt109 inhibitors.

7.2 Experimental Section

7.2.1 Data Sets

The structures and activity information were from the HTS1 experiment, where we evaluated 82,861 compounds from our in-house collection for activity against Rtt109 complex. The detailed assay description can be found in Chapter 2.

In HTS1, a total of 213 compounds (dataset **ITB**) were confirmed via dose-response to be active against Rtt109 complex, with activity ranging from 0.49 – 17.5μM. A total of 40769 compounds (dataset **ITD)** with percentage inhibition of less than 40% were classified as inactives. The 200 compounds and 10528 compounds comprising inactive control dataset **ITC** and **ITE** respectively were chosen from ITD using the Tanimoto-based diverse subset selection in Pipeline Pilot.[123]

7.2.2 Computational Methods

7.2.2.1 3D Structure Generation

Three-dimensional conformations of all dataset structures were generated via geometry optimization by energy minimization in Pipeline Pilot and were further geometry optimized in MOE.

7.2.2.2 Descriptors Calculation

MOE descriptors. Molecular descriptors were used to quantitatively represent structural and physicochemical properties of compounds. A total of 334 both 2D and 3D molecular descriptors were calculated using MOE.[128] These descriptors included physical properties, atom and bond counts, connectivity and shape indices, MOPAC descriptors, partial charge descriptors, pharmacophore feature descriptors, and so on. Details of the descriptors can be found in MOE manual or at http://www.chemcomp.com/journal/descr.htm.

Extended-Connectivity Fingerprints (ECFP). Extended-Connectivity Fingerprints are circular topological fingerprints based on the 2D structure of a compound.[129] The ECFP descriptor generation process systematically collects the neighboring atoms of each

non-hydrogen atom within multiple circular layers up to a given diameter and assigns them integer codes (known as identifiers). Hash functions are then applied to the set of the resulting identifiers to produce the fixed-length binary representation that defines the extended-connectivity fingerprint. A diameter of six was used to calculate the ECFP-6 descriptors for compounds used in our study.

7.2.2.3 Data Set Division for Model Development and Validation

Compounds in ITB and ITC were combined and randomly split into a training set (Train 1) of 330 compounds (170 actives and 160 inactives, 80% of ITB and ITC) and an external test set (Test 1) of 83 compounds (43 actives and 40 inactives, 20% of ITB and ITC).

Because there are many more inactives than actives, ITB with a very limited number of diverse inactives may not be sufficient to represent the chemical space of the inactives. Studies have demonstrated that models built on a limited number of compounds are expected to have limited applicability and may result in unreliable predictions when applied to virtual screening of structurally diverse chemical databases.[130, 131] The significantly larger number of inactives in the training set is expected to increase the applicability domain of the model and to reduce the rate of false positives. Therefore, we decided to include a much larger number of inactives in our training set. Similarly, compounds in ITB and ITE were merged and randomly split into a training set (Train 2) of 8592 compounds (170 actives and 8422 inactives, 80% of ITB and ITE) and an external test set (Test 2) of 2149 compounds which included 43 actives and 2106 inactives (20% of ITB and ITE).

7.2.2.4 Support Vector Machine

In this study, all SVM models were built and optimized using the open source package RapidMiner.[132] We used Gaussian Radial Basis Function as the kernel type：

$$k(x^{(i)}, x^{(j)}) = \exp(-\gamma||x^{(i)} - x^{(j)}||)^2$$

Parameters C and γ were optimized by maximizing the prediction accuracy in a 10-fold cross validation of the training data. Accuracy, sensitivity (recall, true positive rate), specificity (true negative rate), precision (positive predictive value, the number of true positives divided by the sum of true positives and false positives), and f measure (the harmonic mean of precision and recall) were employed to evaluate the prediction power of the model against the external test sets.

7.3 Results and Discussion

7.3.1 Model Performances

Table 7.1 shows the internal predictive performances of the models based on the training set Train 1 by means of 10-fold cross-validation using MOE descriptors. The combination of C = 1.26 and γ = 0.01 yielded the model (Model 1) with the best cross-validated accuracy of 96.97% for Train 1 using MOE descriptors. Model 1 demonstrated its predictive power through the performance on the blind test set Test 1 and the predicting results are displayed in Table 7.2. A total of 40 out of 40 inactives (100%) were accurately classified. Only 3 out of 40 inactives (6.12%) in Test 1 were incorrectly predicted to be inactives.

We then analyzed the compounds that were predicted incorrectly in the 10-fold cross validation process and in the external evaluation of Model 1. Table 7.3 shows the structures of the misclassified compounds in the 10-fold validation process. By analyzing

their structures, it is easy to find that the obvious structural differences between some of these compounds and most of the actives may explain why they were predicted to be inactive. For example, there is almost no positive charged compound in the active set, and thus the positive charged N atom of compound GPHR-00032548 may be the reason why it was predicted to be inactive. Compound GPHR-00029897 contains a fused ring system and compound GPHR-00031010 contains many O atoms. Both structural properties are very rare in compounds in the active set. In the case of compound GPHR-00058762, its size is much smaller than the majority of actives. As for compound GPHR-00026937, it is the only active compound with a 6-member ring in the position connecting to the S atom among the compounds with the succinamide scaffold. Similarly, the misclassified compounds in the Test 1 set shown in Table 7.4 are structurally dissimilar to the most of the actives in Train 1. The size of GPHR-00044631 may be too small. Compound GPHR-00024372 contains several methoxy groups and compound GPHR-00027595 contains several hydroxyl groups, which are not present in the actives in Train 1. More information related to why these compounds are actives may be needed in order to predict these compounds correctly. It is also possible that some compounds are outliers that cannot be directly identified by SVM models. It is an interesting topic for future work.

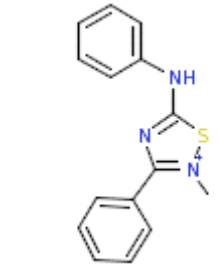
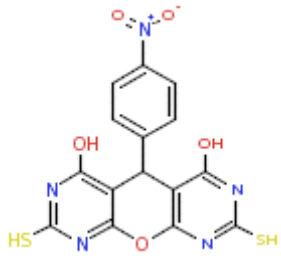
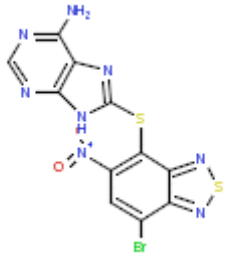
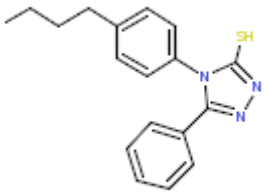
Table 7.1. The performance of the internal 10-fold cross validation of the model (Model 1) developed based on database Train 1 using MOE descriptors

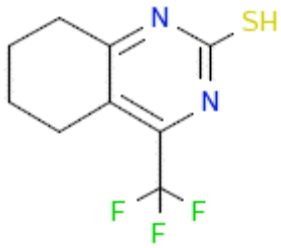
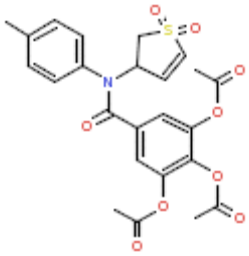
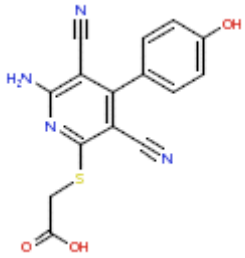
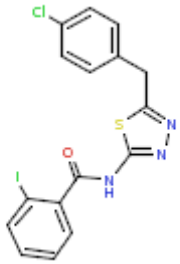
	true active	true inactive	class precision
pred. active	160	0	100%
pred. inactive	10	160	94.12%
class recall	94.12%	100%	
C = 1.26, γ = 0.01, cross-validated accuracy = 96.97%			

Table 7.2. The performance of the model (Model 1) developed based on database Train 1 using MOE descriptors on the external test set Test 1

	true active	true inactive	class precision
pred. active	40	0	100%
pred. inactive	3	40	93.02%
class recall	93.02%	100%	
C = 1.26, γ = 0.01, accuracy = 96.39%			

Table 7.3. The structures of the incorrectly predicted compounds in the 10-fold validation process of Model 1

Structures	Compound ID	IC_{50} (μM)	Prediction
	GPHR-00032548	0.55	inactive
	GPHR-00029897	1.28	inactive
	GPHR-00051573	1.43	inactive
	GPHR-00057665	1.7	inactive

	GPHR-00058762	2.36	inactive
	GPHR-00031010	2.37	inactive
	GPHR-00001819	2.96	inactive
	GPHR-00003628	4.69	inactive

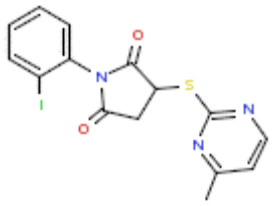
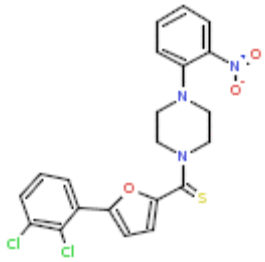
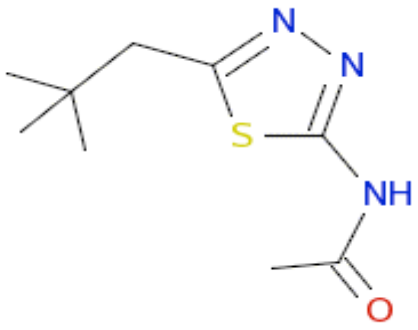
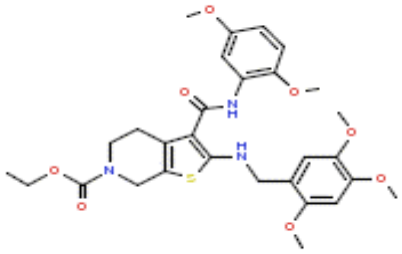
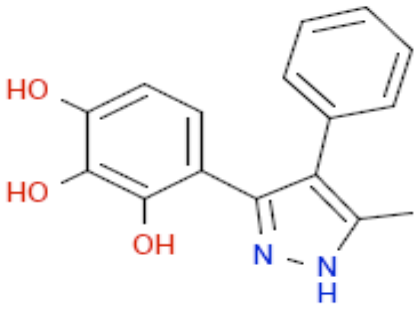
	GPHR-00026937	6.73	inactive
	GPHR-00025780	7.77	inactive

Table 7.4. The structures of incorrectly predicted compounds in the external validation of Model 1 using Test 1

Structures	Compound ID	IC_{50} (μM)	Prediction
	GPHR-00044631	3.25	inactive
	GPHR-00024372	4.41	inactive
	GPHR-00027595	5.74	inactive

A SVM model was also developed using the training set Train 1 with ECFP-6 descriptors. The cross-validated accuracy, sensitivity, and specificity of the optimized model (Model 2) using ECFP-6 descriptors for Train 1 were 92.19%, 92.94%, and 91.25%, respectively. The performance of the internal validation of Model 2 is worse than that of Model 1 and thus ECFP-6 descriptors were not used in the further model developing process. The internal evaluation results are summarized in Table 7.5.

Table 7.5. The performance of the internal 10-fold cross validation of the model (Model 2) developed based on database Train 1 using ECFP-6 descriptors

	true active	true inactive	class precision
pred. active	158	14	91.86%
pred. inactive	12	146	92.41%
class recall	92.94%	91.25%	
C = 10.01, γ = 0.01, cross-validated accuracy = 92.19%			

As mentioned in the section 7.2.2.3, in order to increase the applicability domain of SVM model, Train 2 which contains a much larger number of diverse inactives than Train 1 was used to train the SVM model. However, the significantly larger number of inactives (8422 inactives) compared to the number of actives (170 actives) in Train 2 leads to a highly imbalanced data set. This biased data set will result in an SVM model that is biased towards prediction of the majority class (inactives) and will be poorly predictive for the actives. To resolve this problem, some techniques can be applied to cope with this imbalanced data set. One of the common techniques is to under-sample the majority class (inactives) in the training set. This is basically what we did for Train 1, which includes comparable numbers of actives and inactives. The drawback of this technique is that the reduction in data for model building will lead to loss of information and a limited application domain of the resulting model. Another technique is to increase the cost associated with misclassification of the minority class (actives). During our SVM model development and optimization process based on Train 2, a balanced cost was applied, i.e., a 49 times as much penalty will be paid for the misclassification of an active compound into an inactive compound compared to the misclassification of an inactive compound into an active compound. The performances of the model (Model 3) built on Train 2 using MOE descriptors and the balanced cost are displayed in Table 7.6. Its cross-validated accuracy, sensitivity, and specificity were 90.84%, 79.41%, and 91.07%, respectively. As you may notice, the sensitivity decreased dramatically compared to that of Model 1. This may be due to the fact that more inactives that are structurally similar to the actives were included in the training set. It is very challenging to discriminate

between actives and inactives with subtle structural differences. In order to make correct predictions, information related to why these small differences in the structure result in totally different activities may be needed.

The performance of Model 3 on the external test set Test 2 is shown in Table 7.7. The overall prediction accuracy, sensitivity, and specificity are comparable to the 10-fold validated ones of Model 3. The results indicate that Model3 is able to retrieve the majority (76.74%) of the actives while rejecting 91.36% of the inactives. We then trained the SVM model with all compounds in ITB and ITE to build a comprehensive model (Model 4) for predicting compounds with unknown activities. Based on the 10-fold validation results shown in Table 7.8, Model 4 demonstrated robust predictive power by correctly identifying 84.51% of the actives and 90.40% of the inactives.

Table 7.6. The performance of the internal 10-fold cross validation of the model (Model 3) developed based on database Train 2 using MOE descriptors

	true active	true inactive	class precision
pred. active	135 (TP)	752	15.22%
pred. inactive	35	7670 (TN)	99.55%
class recall	79.41%	91.07%	
C = 1, γ = 0.001, cross-validated accuracy = 90.84%			

Table 7.7. The performance of the model (Model 3) developed based on database Train 2 using MOE descriptors on the external test set Test 2

	true active	true inactive	class precision
pred. active	33 (TP)	192	14.67%
pred. inactive	10	1924 (TN)	99.48%
class recall	76.74%	91.36%	
C = 1, γ = 0.001, cross-validated accuracy = 90.84%			

Table 7.8. The performance of the internal 10-fold cross validation of the model (Model 4) developed based on database ITB and ITE using MOE descriptors

	true active	true inactive	class precision
pred. active	180 (TP)	1011	15.11%
pred. inactive	33	9517 (TN)	99.65%
class recall	84.51%	90.40%	
$C = 0.5$, $\gamma = 0.001$, cross-validated accuracy = 90.28%			

7.3.2 Analysis of Molecular Descriptors

Several molecular descriptors have higher SVM kernel weights than others, suggesting they play more important roles in classifying actives and inactives. These descriptors for Model 4 are shown in Table 7.9 and the corresponding descriptor value distribution for the actives and inactives is displayed in Figure 7.2. We also noticed that the mean values of these descriptors for the actives had significant differences from those of the inactives. In Model 4, descriptor a_nS, the number of S atoms, has a higher mean value for the actives than for the inactives while descriptors a_nCl, the number of Cl atoms, and a_nN, the number of N atoms have lower mean values for the actives than for the inactives. Figure 7.2 illustrates that the majority of inactives do not contain S atoms while most actives contain at least one S atom. As for the number of Cl atoms, although the difference between the mean values for actives and inactives is not large, the value distribution indicates that the inactives are more likely to have more than two Cl atoms than the actives. The value of the descriptor "reactive", which represents the number of reactive groups, for actives is smaller than that for the inactives. The value distribution figures show that none of the actives contain reactive groups while a small portion of inactives contain one reactive group. The value distribution figures also demonstrate that a higher percentage of inactives has one, three, or five N atoms than that of the actives. The PM3_LUMO descriptor shows the energy (eV) of the Lowest Unoccupied Molecular Orbital of compounds calculated using the PM3 method. The lower PM3_LUMO values for the actives indicate the actives are more electrophilic compared to the inactives. The TPSA, polar surface area ($Å^2$), is well correlated with the oral bioavailability of drug

candidates.[133] The average value of TPSA for actives is larger than that for inactives, suggesting that the actives are likely to have more solvent-exposed polar groups. The descriptor b_rotR is a rotatable bond related descriptor, and it has a higher mean value for the inactives than for the actives. It indicates that among the bonds between heavy atoms that the compounds have, there are more acyclic single bonds in inactive compounds than the actives, and the active compounds tend be more rigid compared to the inactives. MOE descriptors PEOE_VSA+2 and PEOE_VSA-1 are the sum of the accessible van der Waals surface area (in $Å^2$) for each atom, of which partial charge is in the range [0.10, 0.15) and [-0.10, -0.05) based on the Partial Equalization of Orbital Electronegativities (PEOE) method.[103] PEOE_VSA+2 has a higher mean value for the actives while the PEOE_VSA-1 has a higher mean value for the inactives. Based on the value distribution in Figure 7.2, most actives and inactives have a PEOE_VSA+2 value ranging from 0 to 20 $Å^2$, but a larger portion of actives has a PEOE_VSA+2 value ranging from 20 to 60 $Å^2$ compared to that of the inactives. The majority of the actives and inactives has a PEOE_VSA-1 value ranging from 0 to 80 $Å^2$, while a larger portion of inactives has a PEOE_VSA-1 value that is larger than 120 $Å^2$. These results suggest that the actives contain more certain types of partially positively charged atoms and the inactives contain more certain types of partially negatively charged atoms. In addition, the average values of the estimated logP values for the actives are slightly larger than the inactives, indicating the actives are likely to be more hydrophobic than the inactives. As shown in Figure 7.2, the majority of both actives and inactives have an estimated logP value

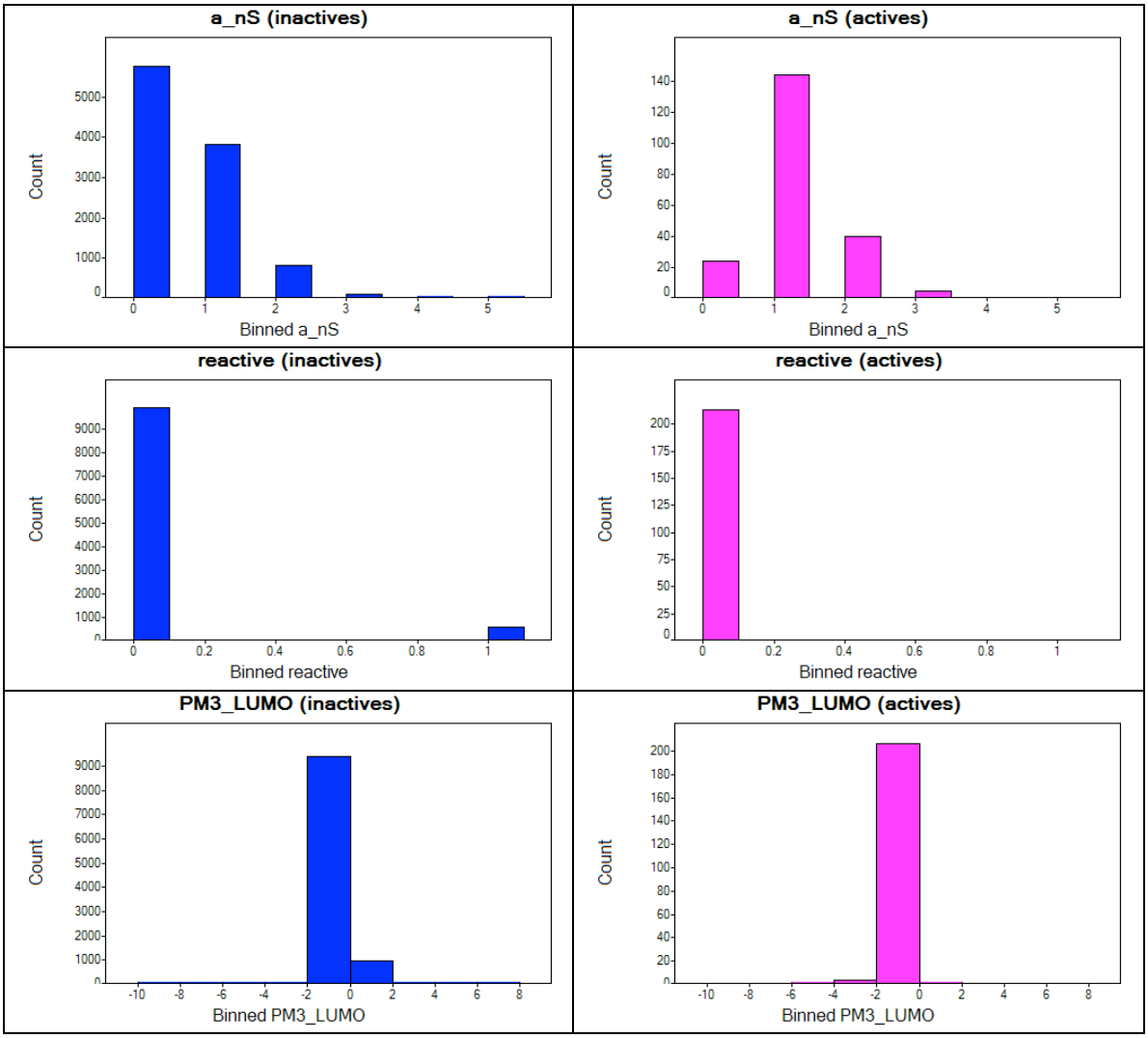
ranging from 2 to 4, but a larger portion of actives has estimated logP values ranging from 4 to 6 than the inactives.

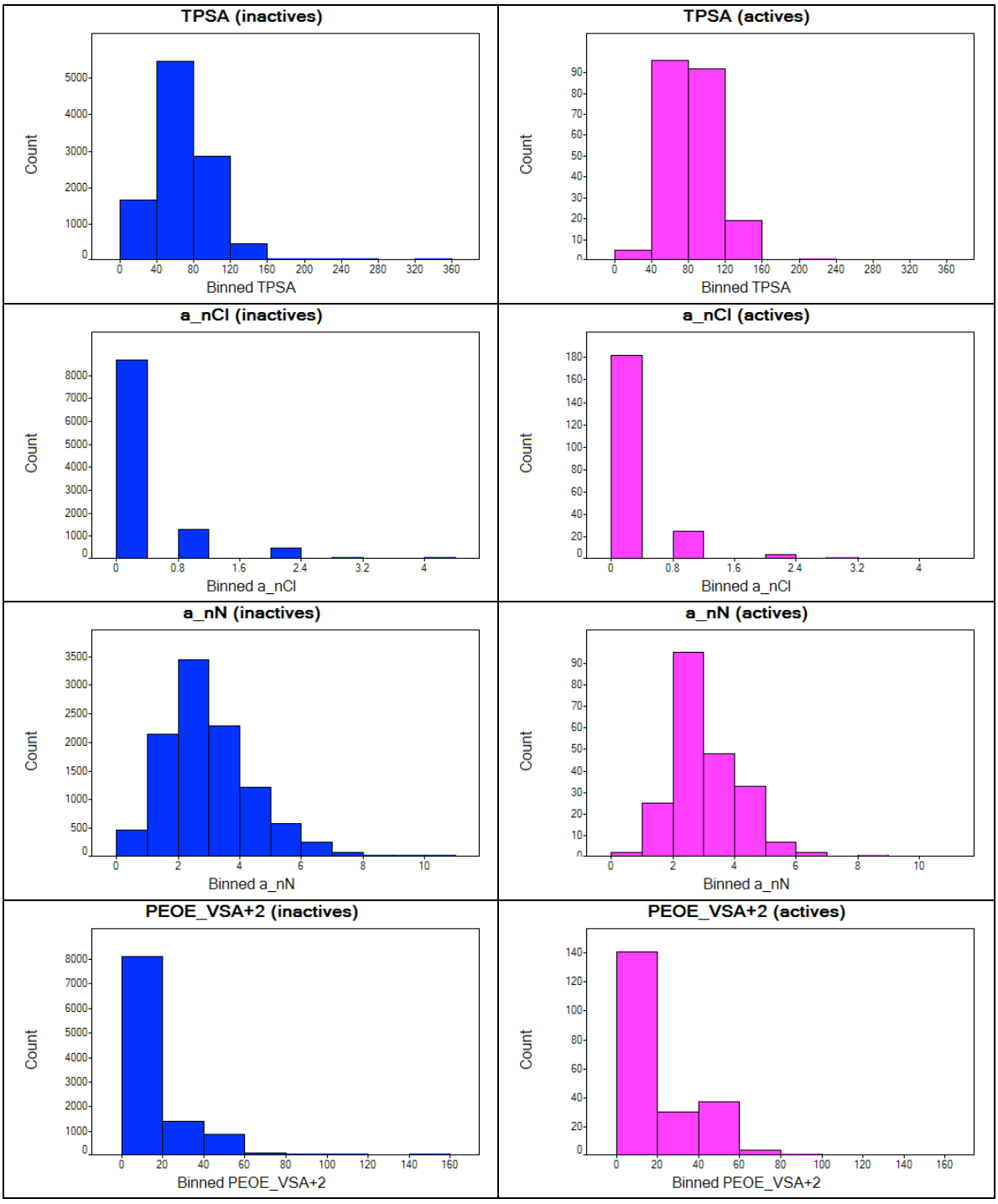
Table 7.9. The representative important MOE descriptors in Model4 and their corresponding average values for actives and inactives in ITB and ITE

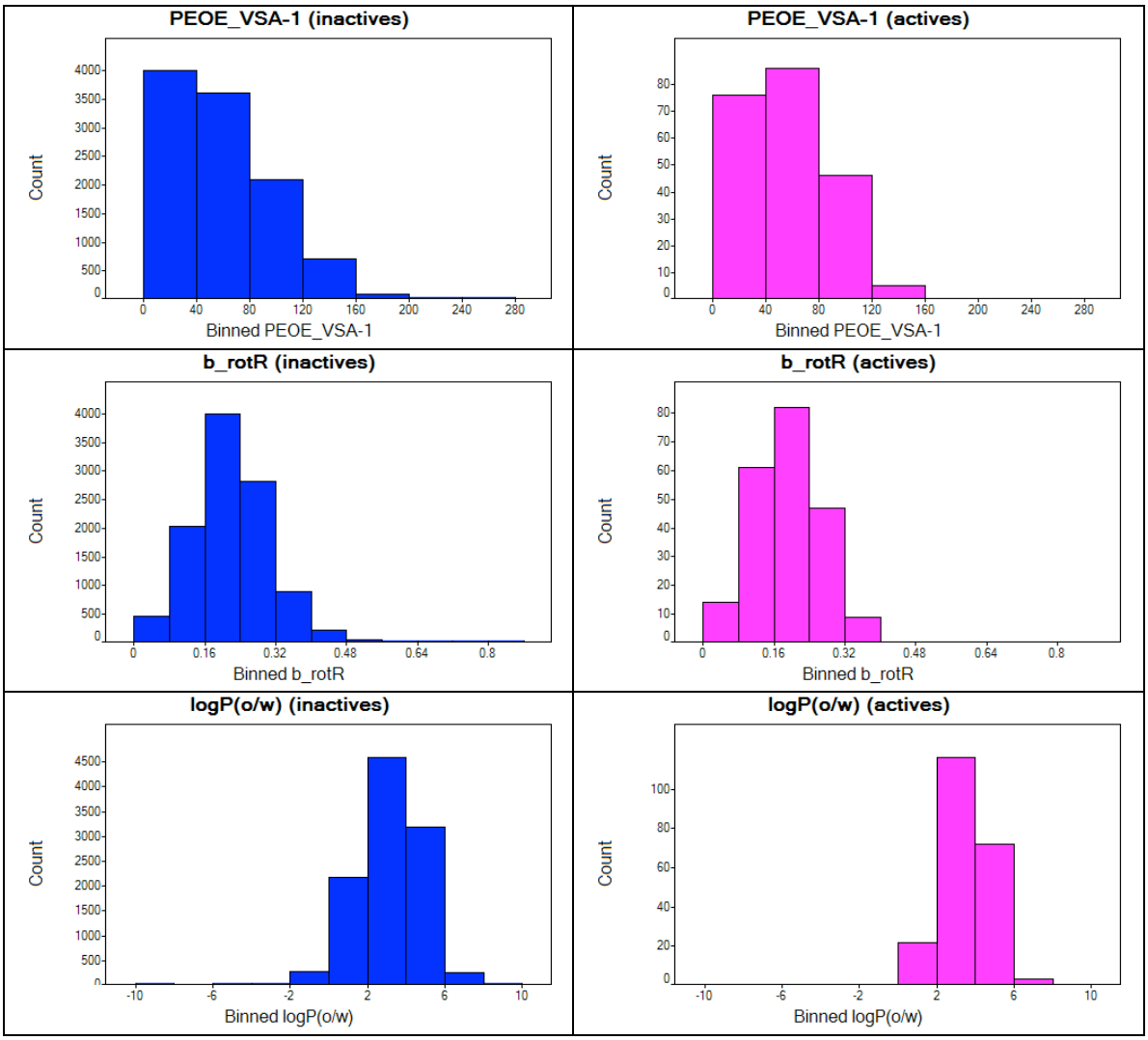
MOE Descriptor	Weight	Average Value (Actives)	Average Value (Inactives)	Meaning
a_nS	157.82	1.12	0.55	Number of sulfur atoms
reactive	-148.05	0.00	0.06	Indicator of the presence of reactive groups in a compound. The definition of reactive groups is based on the Oprea set[134]. The reactive groups include metals, phospho-, N/O/S-N/O/S single bonds, thiols, acyl halides, Michael Acceptors, azides, and esters.
PM3_LUMO	-124.85	-1.15	-0.79	The energy (eV) of the Lowest Unoccupied Molecular Orbital (LUMO) calculated using the PM3.[135-137]
TPSA	123.42	83.41	69.09	Polar surface area ($Å^2$) calculated using group contributions to approximate the polar surface area from connection table information only and using the method of Ertl et al.[138]
a_nCl	-104.52	0.18	0.24	Number of chlorine atoms
a_nN	-101.74	2.47	2.56	Number of nitrogen atoms
PEOE_VSA+2	95.00	17.92	12.80	The sum of van der Waals surface area (in $Å^2$) for each atom, of which partial charge is in the range [0.10, 0.15) based on the Partial Equalization of Orbital Electronegativities (PEOE) method.[103]
PEOE_VSA-1	-88.83	54.61	57.74	The sum of van der Waals surface area (in $Å^2$) for each atom, of which partial charge is in the range [-0.10, -0.05) based on the Partial Equalization of Orbital Electronegativities (PEOE) method.[103]

b_rotR	-81.27	0.19	0.22	Fraction of rotatable bonds, which is defined as the number of rotatable bonds divided by the number of bonds between heavy atoms.
logP(o/w)	76.08	3.58	3.17	Log of the octanol/water partition coefficient (including implicit hydrogens) calculated from a linear atom type model with $r^2 = 0.931$, RMSE=0.393 that was trained using 1,827 molecules.[128]

Figure 7.2. The representative important MOE descriptor property distribution profiles for 213 Rtt109-complex inhibitors (magenta, right), compared to 10528 inactives (blue, left)



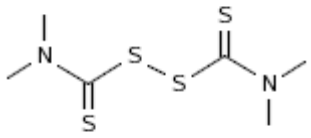
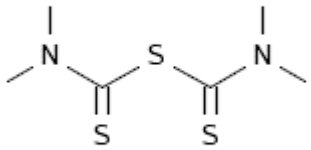
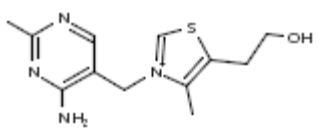

TPSA (inactives)
Count
Binned TPSA
TPSA (actives)
Count
Binned TPSA
a_nCl (inactives)
Count
Binned a_nCl
a_nCl (actives)
Count
Binned a_nCl
a_nN (inactives)
Count
Binned a_nN
a_nN (actives)
Count
Binned a_nN
PEOE_VSA+2 (inactives)
Count
Binned PEOE_VSA+2
PEOE_VSA+2 (actives)
Count
Binned PEOE_VSA+2

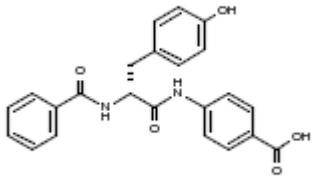
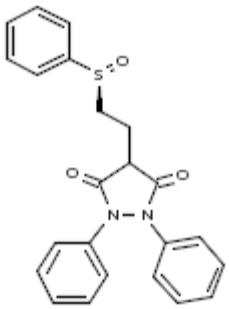
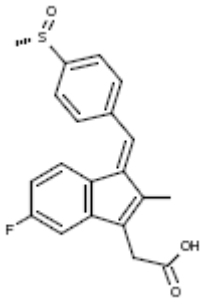

PEOE_VSA-1 (inactives)
Count
Binned PEOE_VSA-1
PEOE_VSA-1 (actives)
Count
Binned PEOE_VSA-1
b_rotR (inactives)
Count
Binned b_rotR
b_rotR (actives)
Count
Binned b_rotR
logP(o/w) (inactives)
Count
Binned logP(o/w)
logP(o/w) (actives)
Count
Binned logP(o/w)

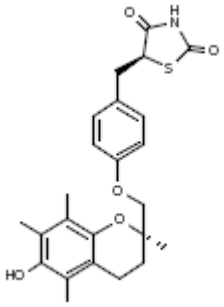
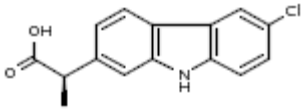
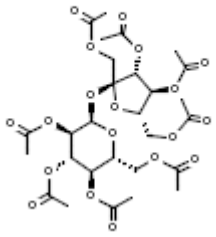
7.3.3 Application of Model for the Virtual Screening to Identify Novel Rtt109 Inhibitors

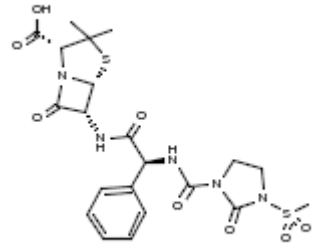
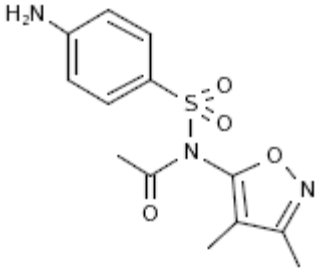
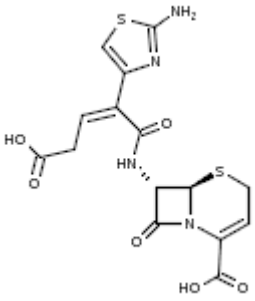
The comprehensive model (Model 4) built based on the database ITB and ITE should be useful for novel Rtt109 inhibitor discovery because of the model's robust internal predictive power. We applied this model to screen the virtual database of FDA approved drugs for humans. The top 20 compounds with a prediction confidence larger than 0.6 are listed in Table 7.10. One obvious property these compounds share with the actives is that they contain at least one S atom. In addition, some of these compounds have the hydrophobic/aromatic and acceptor features that were pinpointed by the pharmacophore models, indicating consistency between the SVM and pharmacophore models. Notably, some of these predicted actives were reported as antibiotics and one of them has antifungal effects. If these compounds have Rtt109 inhibitory activities in the experimental assay, they may be able to be used as dual drugs with both antibiotic and antifungal effects or as drugs with different antifungal mechanisms.

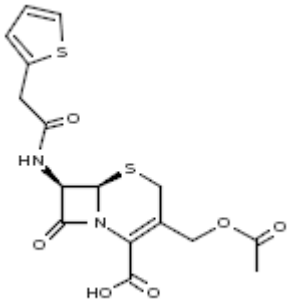
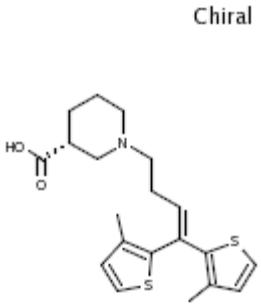
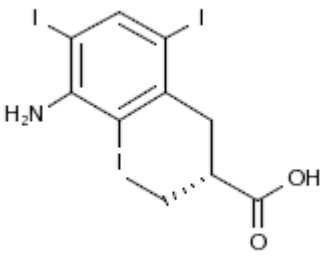
Table 7.10. The top 20 compounds with a prediction confidence larger than 0.6 in the database of FDA approved drugs for humans

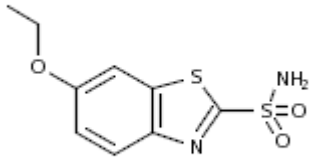
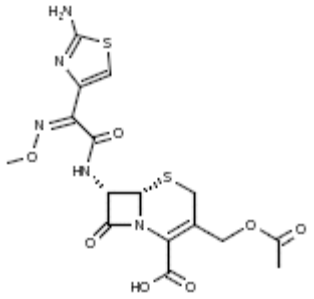
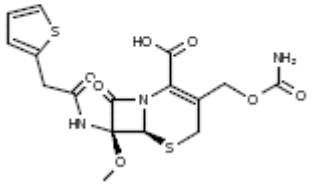
Structures	Name	Drug Usage	Confidence (active)
	Nobecutan	Antiseptic	0.81
	Tetramethylthiuram monosulfide	Antifungal	0.76
	Thiaminum	Vitamin	0.72

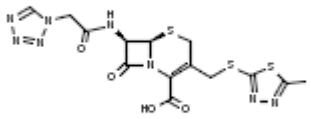
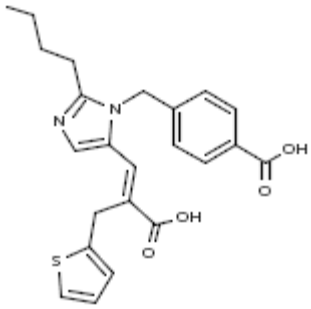
	N-benzoyl-l-tyrosyl-paba	Diagnostic Aid	0.70
	Sulfinpyrazone	Antiurolithic	0.70
	Sulindac	Antineoplastic	0.68

	Troglitazone	Antidiabetic	0.68
	Carprofen	Anti- inflammatory	0.67
Chiral 	Sucrose octaacetate, ((alpha-d)- fructofuranosyl)-isomer	Phamaceutic Aid	0.66

	Mezlocillin	Antibiotic	0.65
	Sulfisoxazole acetal	Antibacterial	0.64
	Ceftibuten	Antibiotic	0.63

	Cephalothin	Antibiotic	0.63
	Tiagabine	Anticonvulsant	0.63
	Iopanoic acid	Diagnostic Aid	0.62

	Ethoxazolamide	Diuretic	0.62
	Cefotaxime	Antibiotic	0.62
Chiral 	Cefoxitin	Antibiotic	0.62

	Cefazolin	Antibiotic	0.61
	Eprosartan	Antihypertensive	0.61

7.4 Concluding Remarks

Different SVM models suitable for virtual screening purposes without the knowledge of 3D information of the protein target were developed from Rtt109 inhibitors with experimental biological activity data. These models were validated using internal validation with 10-fold cross-validation and external validation with compounds not used in the model development. A model (Model 1) with the MOE descriptors based on the randomly split balanced training set yielded the highest cross-validated accuracy, 96.97% on the internal test set. It achieved an accuracy of 96.39% on a heterogeneous external test set. A comprehensive model (Model 4) with the MOE descriptors based on 213 actives and 10528 inactives was developed in order to expand the application domain of the model. Its 10-fold cross-validated accuracy was 90.28%. It is expected to retrieve 84.51% actives and reject 90.40% inactives based on its performance in the internal validation process.

Several molecular descriptors including the number of S atoms, Cl atoms, N atoms, reactive groups, the energy of the Lowest Unoccupied Molecular Orbital, polar surface area, partial charge distribution, rotatable bonds, and logP(o/w), may play a critical role in defining Rtt109inhibitory activities. These important descriptors indicated specific atomic, partial charge, and other property requirements for Rtt109inhibitors.

In summary, we have developed internally validated and externally predictive SVM models for Rtt109 inhibitory activities based on HTS1. Although utilizing different methodology, SVM, 3D-QSAR, and pharmacophore models all were able to accurately predict activities of compounds and agreed with each other. These models presented in

this work are of potential use to complement HTS in screening chemical libraries and identifying novel Rtt109 inhibitors. However, further follow-up experiments may indicate some of the actives used in the model development may be false positives due to their interference with the assay reagents. In future work, further validated experimental activity data will be needed to verify and refine SVM models.

REFERENCES

1. Fisher, M. Emerging fungal threats to animal, plant and ecosystem health. *Mycoses* **2012,** 55, 79-80.
2. Giraud, T.; Gladieux, P.; Gavrilets, S. Linking the emergence of fungal plant diseases with ecological speciation. *Trends Ecol. Evol.* **2010,** 25, 387-395.
3. Gargas, A.; Trest, M. T.; Christensen, M.; Volk, T. J.; Blehert, D. S. Geomyces destructans sp nov associated with bat white-nose syndrome. *Mycotaxon* **2009,** 108, 147-154.
4. http://news.dnr.state.mn.us/2013/08/09/fungus-dangerous-to-bats-detected-at-2-minnesota-state-parks/#more-12787 (December 1st, 2013).
5. Crawford, A. J.; Lips, K. R.; Bermingham, E. Epidemic disease decimates amphibian abundance, species diversity, and evolutionary history in the highlands of central Panama. *Proc. Natl. Acad. Sci. U. S. A.* **2010,** 107, 13777-13782.
6. Clark, T. A.; Hajjeh, R. A. Recent trends in the epidemiology of invasive mycoses. *Curr. Opin. Infect. Dis.* **2002,** 15, 569-574.
7. Fleming, R. V.; Walsh, T. J.; Anaissie, E. J. Emerging and less common fungal pathogens. *Infect. Dis. Clin. North Am.* **2002,** 16, 915-933.
8. Romani, L. Immunity to fungal infections. *Nat. Rev. Immunol.* **2011,** 11, 275-288.
9. Havlickova, B.; Czaika, V. A.; Friedrich, M. Epidemiological trends in skin mycoses worldwide. *Mycoses* **2008,** 51, 2-15.
10. Marr, K. A.; Carter, R. A.; Crippa, F.; Wald, A.; Corey, L. Epidemiology and outcome of mould infections in hematopoietic stem cell transplant recipients. *Clin. Infect. Dis.* **2002,** 34, 909-917.
11. Dagenais, T. R. T.; Keller, N. P. Pathogenesis of Aspergillus fumigatus in Invasive Aspergillosis. *Clin. Microbiol. Rev.* **2009,** 22, 447-465.
12. Pfaller, M. A.; Diekema, D. J. Epidemiology of invasive candidiasis: a persistent public health problem. *Clin. Microbiol. Rev.* **2007,** 20, 133-163.
13. Petrikkos, G.; Skiada, A. Recent advances in antifungal chemotherapy. *Int. J. Antimicrob. Agents* **2007,** 30, 108-117.
14. Odds, F. C.; Brown, A. J. P.; Gow, N. A. R. Antifungal agents: mechanisms of action. *Trends Microbiol.* **2003,** 11, 272-279.
15. Saliba, F.; Dupont, B. Renal impairment and Amphotericin B formulations in patients with invasive fungal infections. *Med. Mycol.* **2008,** 46, 97-112.
16. Anderson, J. B. Evolution of antifungal-drug resistance: Mechanisms and pathogen fitness. *Nat. Rev. Microbiol.* **2005,** 3, 547-556.
17. Klepser, M. E. Candida resistance and its clinical relevance. *Pharmacotherapy* **2006,** 26, 68s-75s.
18. Espinel-Ingroff, A. Mechanisms of resistance to antifungal agents: Yeasts and filamentous fungi. *Rev. Iberoam. Micol.* **2008,** 25, 101-106.
19. Brown, G. D.; Denning, D. W.; Gow, N. A. R.; Levitz, S. M.; Netea, M. G.; White, T. C. Hidden Killers: Human Fungal Infections. *Sci. Transl. Med.* **2012,** 4.

20. Luger, K.; Mader, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. *Nature* **1997,** 389, 251-260.
21. Kornberg, R. D.; Lorch, Y. L. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **1999,** 98, 285-294.
22. Kouzarides, T. Chromatin modifications and their function. *Cell* **2007,** 128, 693-705.
23. Goll, M. G.; Bestor, T. H. Histone modification and replacement in chromatin activation. *Genes Dev.* **2002,** 16, 1739-1742.
24. Furdas, S. D.; Kannan, S.; Sippl, W.; Jung, M. Small Molecule Inhibitors of Histone Acetyltransferases as Epigenetic Tools and Drug Candidates. *Arch. Pharm. (Weinheim)* **2012,** 345, 7-21.
25. Keppler, B. R.; Archer, T. K. Chromatin-modifying enzymes as therapeutic targets - Part 1. *Expert Opin. Ther. Targets* **2008,** 12, 1301-1312.
26. Kimura, A.; Matsubara, K.; Horikoshi, M. A decade of histone acetylation: Marking eukaryotic chromosomes with specific codes. *J. Biochem. (Tokyo)* **2005,** 138, 647-662.
27. Lee, K. K.; Workman, J. L. Histone acetyltransferase complexes: one size doesn't fit all. *Nat. Rev. Mol. Cell Biol.* **2007,** 8, 284-295.
28. Driscoll, R.; Hudson, A.; Jackson, S. P. Yeast Rtt109 promotes genome stability by acetylating histone H3 on lysine 56. *Science* **2007,** 315, 649-652.
29. da Rosa, J. L.; Boyartchuk, V. L.; Zhu, L. J.; Kaufman, P. D. Histone acetyltransferase Rtt109 is required for Candida albicans pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **2010,** 107, 1594-1599.
30. Wurtele, H.; Tsao, S.; Lepine, G.; Mullick, A.; Tremblay, J.; Drogaris, P.; Lee, E. H.; Thibault, P.; Verreault, A.; Raymond, M. Modulation of histone H3 lysine 56 acetylation as an antifungal therapeutic strategy. *Nat. Med.* **2010,** 16, 774-U73.
31. Tang, Y.; Holbert, M. A.; Wurtele, H.; Meeth, K.; Rocha, W.; Gharib, M.; Jiang, E.; Thibault, P.; Verreault, A.; Cole, P. A.; Marmorstein, R. Fungal Rtt109 histone acetyltransferase is an unexpected structural homolog of metazoan p300/CBP (vol 15, pg 738, 2008). *Nat. Struct. Mol. Biol.* **2008,** 15, 998-998.
32. Han, J. H.; Zhou, H.; Li, Z. H.; Xu, R. M.; Zhang, Z. G. Acetylation of lysine 56 of histone H3 catalyzed by RTT109 and regulated by ASF1 is required for replisome integrity. *J. Biol. Chem.* **2007,** 282, 28587-28596.
33. Su, D.; Hu, Q.; Zhou, H.; Thompson, J. R.; Xu, R. M.; Zhang, Z. G.; Mer, G. Structure and Histone Binding Properties of the Vps75-Rtt109 Chaperone-Lysine Acetyltransferase Complex. *J. Biol. Chem.* **2011,** 286.
34. Kolonko, E. M.; Albaugh, B. N.; Lindner, S. E.; Chen, Y.; Satyshur, K. A.; Arnold, K. M.; Kaufman, P. D.; Keck, J. L.; Denu, J. M. Catalytic activation of histone acetyltransferase Rtt109 by a histone chaperone. *Proc. Natl. Acad. Sci. U. S. A.* **2010,** 107, 20275-80.
35. Albaugh, B. N.; Kolonko, E. M.; Denu, J. M. Kinetic Mechanism of the Rtt109-Vps75 Histone Acetyltransferase-Chaperone Complex. *Biochemistry* **2010,** 49, 6375-6385.

36. Berndsen, C. E.; Tsubota, T.; Lindner, S. E.; Lee, S.; Holton, J. M.; Kaufman, P. D.; Keck, J. L.; Denu, J. M. Molecular functions of the histone acetyltransferase chaperone complex Rtt109-Vps75. *Nat. Struct. Mol. Biol.* **2008,** 15, 948-56.
37. Tsubota, T.; Berndsen, C. E.; Erkmann, J. A.; Smith, C. L.; Yang, L. H.; Freitas, M. A.; Denu, J. M.; Kaufman, P. D. Histone H3-K56 acetylation is catalyzed by histone chaperone-dependent complexes. *Mol. Cell* **2007,** 25, 703-712.
38. Stavropoulos, P.; Nagy, V.; Blobel, G.; Hoelz, A. Molecular basis for the autoregulation of the protein acetyl transferase Rtt109. *Proc. Natl. Acad. Sci. U. S. A.* **2008,** 105, 12236-12241.
39. Albaugh, B. N.; Arnold, K. M.; Lee, S.; Denu, J. M. Autoacetylation of the Histone Acetyltransferase Rtt109. *J. Biol. Chem.* **2011,** 286, 24694-24701.
40. Wang, L.; Tang, Y.; Cole, P. A.; Marmorstein, R. Structure and chemistry of the p300/CBP and Rtt109 histone acetyltransferases: implications for histone acetyltransferase evolution and function. *Curr. Opin. Struct. Biol.* **2008,** 18, 741-747.
41. Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **2010,** 10, 95-115.
42. Liu, X.; Wang, L.; Zhao, K. H.; Thompson, P. R.; Hwang, Y.; Marmorstein, R.; Cole, P. A. The structural basis of protein acetylation by the p300/CBP transcriptional coactivator. *Nature* **2008,** 451, 846-850.
43. Fillingham, J.; Recht, J.; Silva, A. C.; Suter, B.; Emili, A.; Stagljar, I.; Krogan, N. J.; Allis, C. D.; Keogh, M. C.; Greenblatt, J. F. Chaperone control of the activity and specificity of the histone H3 acetyltransferase Rtt109. *Mol. Cell. Biol.* **2008,** 28, 4342-4353.
44. Lin, C. Q.; Yuan, Y. A. Structural Insights into Histone H3 Lysine 56 Acetylation by Rtt109. *Structure* **2008,** 16, 1503-1510.
45. The PyMOL Molecular Graphics System, Version 1.4, Schrödinger, LLC.
46. Buczek-Thomas, J. A.; Hsia, E.; Rich, C. B.; Foster, J. A.; Nugent, M. A. Inhibition of histone acetyltransferase by glycosaminoglycans. *J. Cell. Biochem.* **2008,** 105, 108-120.
47. Eisenberg, T.; Knauer, H.; Schauer, A.; Buttner, S.; Ruckenstuhl, C.; Carmona-Gutierrez, D.; Ring, J.; Schroeder, S.; Magnes, C.; Antonacci, L.; Fussi, H.; Deszcz, L.; Hartl, R.; Schraml, E.; Criollo, A.; Megalou, E.; Weiskopf, D.; Laun, P.; Heeren, G.; Breitenbach, M.; Grubeck-Loebenstein, B.; Herker, E.; Fahrenkrog, B.; Frohlich, K. U.; Sinner, F.; Tavernarakis, N.; Minois, N.; Kroemer, G.; Madeo, F. Induction of autophagy by spermidine promotes longevity. *Nat. Cell Biol.* **2009,** 11, 1305-U102.
48. Balasubramanyam, K.; Swaminathan, V.; Ranganathan, A.; Kundu, T. K. Small molecule modulators of histone acetyltransferase p300. *J. Biol. Chem.* **2003,** 278, 19134-19140.
49. Sun, Y. L.; Jiang, X. F.; Chen, S. J.; Price, B. D. Inhibition of histone acetyltransferase activity by anacardic acid sensitizes tumor cells to ionizing radiation. *FEBS Lett.* **2006,** 580, 4353-4356.
50. Balasubramanyam, K.; Varier, R. A.; Altaf, M.; Swaminathan, V.; Siddappa, N. B.; Ranga, U.; Kundu, T. K. Curcumin, a novel p300/CREB-binding protein-specific inhibitor of acetyltransferase, represses the acetylation of histone/nonhistone proteins and

histone acetyltransferase-dependent chromatin transcription. *J. Biol. Chem.* **2004,** 279, 51163-51171.
51. Hsu, C. H.; Cheng, A. L. Clinical studies with curcumin. *Molecular Targets and Therapeutic Uses of Curcumin in Health and Disease* **2007,** 595, 471-480.
52. Balasubramanyam, K.; Altaf, M.; Varier, R. A.; Swaminathan, V.; Ravindran, A.; Sadhale, P. P.; Kundu, T. K. Polyisoprenylated benzophenone, garcinol, a natural histone acetyltransferase inhibitor, represses chromatin transcription and alters global gene expression. *J. Biol. Chem.* **2004,** 279, 33716-33726.
53. Arif, M.; Pradhan, S. K.; Thanuia, G. R.; Vedamurthy, B. M.; Agrawal, S.; Dasgupta, D.; Kundu, T. K. Mechanism of p300 Specific Histone Acetyltransferase Inhibition by Small Molecules. *J. Med. Chem.* **2009,** 52, 267-277.
54. Choi, K. C.; Jung, M. G.; Lee, Y. H.; Yoon, J. C.; Kwon, S. H.; Kang, H. B.; Kim, M. J.; Cha, J. H.; Kim, Y. J.; Jun, W. J.; Lee, J. M.; Yoon, H. G. Epigallocatechin-3-Gallate, a Histone Acetyltransferase Inhibitor, Inhibits EBV-Induced B Lymphocyte Transformation via Suppression of ReIA Acetylation. *Cancer Res.* **2009,** 69, 583-592.
55. Ravindra, K. C.; Selvi, B. R.; Arif, M.; Reddy, B. A. A.; Thanuja, G. R.; Agrawal, S.; Pradhan, S. K.; Nagashayana, N.; Dasgupta, D.; Kundu, T. K. Inhibition of Lysine Acetyltransferase KAT3B/p300 Activity by a Naturally Occurring Hydroxynaphthoquinone, Plumbagin. *J. Biol. Chem.* **2009,** 284, 24453-24464.
56. Lau, O. D.; Kundu, T. K.; Soccio, R. E.; Ait-Si-Ali, S.; Khalil, E. M.; Vassilev, A.; Wolffe, A. P.; Nakatani, Y.; Roeder, R. G.; Cole, P. A. HATs off: Selective synthetic inhibitors of the histone acetyltransferases p300 and PCAF. *Mol. Cell* **2000,** 5, 589-595.
57. Wu, J.; Xie, N.; Wu, Z.; Zhang, Y.; Zheng, Y. G. Bisubstrate Inhibitors of the MYST HATs Esa1 and Tip60. *Bioorg. Med. Chem.* **2009,** 17, 1381-6.
58. Bandyopadhyay, K.; Baneres, J. L.; Martin, A.; Blonski, C.; Parello, J.; Gjerset, R. A. Spermidinyl-CoA-based HAT inhibitors block DNA repair and provide cancer-specific chemo-and radiosensitization. *Cell Cycle* **2009,** 8, 2779-2788.
59. Biel, M.; Kretsovali, A.; Karatzali, E.; Papamatheakis, J.; Giannis, A. Design, synthesis, and biological evaluation of a small-molecule inhibitor of the histone acetyltransferase Gcn5. *Angewandte Chemie-International Edition* **2004,** 43, 3974-3976.
60. Mai, A.; Rotili, D.; Tarantino, D.; Ornaghi, P.; Tosi, F.; Vicidomini, C.; Sbardella, G.; Nebbioso, A.; Miceli, M.; Altucci, L.; Filetici, P. Small-molecule inhibitors of histone acetyltransferase activity: Identification and biological properties. *J. Med. Chem.* **2006,** 49, 6897-6907.
61. Stimson, L.; Rowlands, M. G.; Newbatt, Y. M.; Smith, N. F.; Raynaud, F. I.; Rogers, P.; Bavetsias, V.; Gorsuch, S.; Jarman, M.; Bannister, A.; Kouzarides, T.; McDonald, E.; Workman, P.; Aherne, G. W. Isothiazolones as inhibitors of PCAF and p300 histone acetyltransferase activity. *Mol. Cancer Ther.* **2005,** 4, 1521-1532.
62. Bowers, E. M.; Yan, G.; Mukherjee, C.; Orry, A.; Wang, L.; Holbert, M. A.; Crump, N. T.; Hazzalin, C. A.; Liszczak, G.; Yuan, H.; Larocca, C.; Saldanha, S. A.; Abagyan, R.; Sun, Y.; Meyers, D. J.; Marmorstein, R.; Mahadevan, L. C.; Alani, R. M.; Cole, P. A. Virtual Ligand Screening of the p300/CBP Histone Acetyltransferase: Identification of a Selective Small Molecule Inhibitor. *Chem. Biol.* **2010,** 17, 471-482.

63. da Rosa, J. L.; Bajaj, V.; Spoonamore, J.; Kaufman, P. D. A small molecule inhibitor of fungal histone acetyltransferase Rtt109. *Bioorg. Med. Chem. Lett.* **2013,** 23, 2853-2859.
64. Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** 103, 11473-11478.
65. Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discovery* **2011,** 10, 188-195.
66. Trievel, R. C.; Li, F. Y.; Marmorstein, R. Application of a fluorescent histone acetyltransferase assay to probe the substrate specificity of the human p300/CBP-associated factor. *Anal. Biochem.* **2000,** 287, 319-328.
67. Chung, C. C.; Ohwaki, K.; Schneeweis, J. E.; Stec, E.; Varnerin, J. P.; Goudreau, P. N.; Chang, A.; Cassaday, J.; Yang, L. H.; Yamakawa, T.; Kornienko, O.; Hodder, P.; Inglese, J.; Ferrer, M.; Strulovici, B.; Kusunoki, J.; Tota, M. R.; Takagi, T. A fluorescence-based thiol quantification assay for ultra-high-throughput screening for inhibitors of coenzyme A production. *Assay Drug Dev. Technol.* **2008,** 6, 361-374.
68. Dahlin, J. L.; Sinville, R.; Solberg, J.; Zhou, H.; Han, J.; Francis, S.; Strasser, J. M.; John, K.; Hook, D. J.; Walters, M. A.; Zhang, Z. A cell-free fluorometric high-throughput screen for inhibitors of rtt109-catalyzed histone acetylation. *PLoS One* **2013,** 8, e78877.
69. Feng, B. Y.; Shoichet, B. K. A detergent-based assay for the detection of promiscuous inhibitors. *Nat. Protoc.* **2006,** 1, 550-553.
70. Ryan, A. J.; Gray, N. M.; Lowe, P. N.; Chung, C. W. Effect of detergent on "promiscuous" inhibitors. *J. Med. Chem.* **2003,** 46, 3448-3451.
71. Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discovery* **2003,** 2, 527-541.
72. Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002,** 9, 646-652.
73. McNaught, A. D.; Wilkinson, A.; International Union of Pure and Applied Chemistry. *Compendium of chemical terminology : IUPAC recommendations*. 2nd ed.; Blackwell Science: Oxford England ; Malden, MA, USA, 1997; p vii, 450 p.
74. Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H. H.; McCammon, J. A. Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **2004,** 47, 1879-1881.
75. Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domoal, R. A.; Wang, L. G.; Anderson, K. S.; Jorgensen, W. L. FEP-guided selection of bicyclic heterocycles in lead optimization for non-nucleoside inhibitors of HIV-1 reverse transcriptase. *J. Am. Chem. Soc.* **2006,** 128, 15372-15373.
76. Amaro, R. E.; Baron, R.; McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput-Aided. Mol. Des.* **2008,** 22, 693-705.

77. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005,** 26, 1781-1802.
78. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010,** 31, 671-690.
79. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998,** 102, 3586-3616.
80. Reichert, D.; Zinkevich, T.; Saalwachter, K.; Krushelnitsky, A. The relation of the X-ray B-factor to protein dynamics: insights from recent dynamic solid-state NMR data. *J. Biomol. Struct. Dyn.* **2012,** 30, 617-627.
81. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery* **2004,** 3, 935-949.
82. Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.* **1998,** 27, 249-284.
83. Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.* **2007,** 21, 281-306.
84. Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003,** 46, 499-511.
85. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006,** 49, 6177-6196.
86. 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.
87. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; 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.
88. Ruppert, J.; Welch, W.; Jain, A. N. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997,** 6, 524-533.
89. Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996,** 3, 449-462.

90. Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput-Aided. Mol. Des.* **1996,** 10, 427-440.
91. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996,** 118, 11225-11236.
92. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions .1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput-Aided. Mol. Des.* **1997,** 11, 425-445.
93. Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008,** 36, D901-D906.
94. http://www.leadquest.com. (Augest 6, 2008).
95. Stevenson, J. M.; Mulready, P. D. Pipeline pilot 2.1. *J. Am. Chem. Soc.* **2003,** 125, 1437-1438.
96. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery. Rev.* **1997,** 23, 3-25.
97. Cramer, R. D.; Wendt, B. Pushing the boundaries of 3D-QSAR. *J. Comput-Aided. Mol. Des.* **2007,** 21, 23-32.
98. Amin, E. A.; Welsh, W. J. Highly predictive CoMFA and CoMSIA models for two series of stromelysin-1 (MMP-3) inhibitors elucidate S1 ' and S1-S2 ' binding modes. *J. Chem. Inf. Model.* **2006,** 46, 1775-1783.
99. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular-Field Analysis (Comfa) .1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988,** 110, 5959-5967.
100. Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indexes in a Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity. *J. Med. Chem.* **1994,** 37, 4130-4146.
101. Vinter, J. G.; Davis, A.; Saunders, M. R. Strategic approaches to drug design. I. An integrated software framework for molecular modelling. *J. Comput-Aided. Mol. Des.* **1987,** 1, 31-51.
102. The Tripos Bookshelf Version X-1.3. In Tripos, I. S. L., MO, 2011.
103. Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges. *Tetrahedron* **1980,** 36, 3219-3228.
104. Purcell, W.; Singer, J. A brief review and table of semiempirical parameters used in the Hueckel molecular orbital method. *J. Chem. Eng. Data.* **1967,** 12, 235-246.
105. Halgren, T. A. Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996,** 17, 490-519.
106. Frank, I. E.; Feikema, J.; Constantine, N.; Kowalski, B. R. Prediction of Product Quality from Spectral Data Using the Partial Least-Squares Method. *J. Chem. Inf. Comput. Sci.* **1984,** 24, 20-24.

107. Heiden, W.; Moeckel, G.; Brickmann, J. A New Approach to Analysis and Display of Local Lipophilicity Hydrophilicity Mapped on Molecular-Surfaces. *J. Comput-Aided. Mol. Des.* **1993,** 7, 503-514.
108. Exner, T.; Keil, M.; Moeckel, G.; Brickmann, J. Identification of substrate channels and protein cavities. *Journal of Molecular Modeling* **1998,** 4, 340-343.
109. Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010,** 53, 539-558.
110. Yang, H. Y.; Shen, Y.; Chen, J. H.; Jiang, Q. F.; Leng, Y.; Shen, J. H. Structure-based virtual screening for identification of novel 11 beta-HSD1 inhibitors. *Eur. J. Med. Chem.* **2009,** 44, 1167-1171.
111. Ryu, K.; Kim, N. D.; Il Choi, S.; Han, C. K.; Yoon, J. H.; No, K. T.; Kim, K. H.; Seong, B. L. Identification of novel inhibitors of HCV RNA-dependent RNA polymerase by pharmacophore-based virtual screening and in vitro evaluation. *Bioorg. Med. Chem.* **2009,** 17, 2975-2982.
112. Taha, M. O.; Dahabiyeh, L. A.; Bustanji, Y.; Zalloum, H.; Saleh, S. Combining Ligand-Based Pharmacophore Modeling, Quantitative Structure-Activity Relationship Analysis and in Silico Screening for the Discovery of New Potent Hormone Sensitive Lipase Inhibitors. *J. Med. Chem.* **2008,** 51, 6478-6494.
113. Chiu, T. L.; Amin, E. A. Development of a Comprehensive, Validated Pharmacophore Hypothesis for Anthrax Toxin Lethal Factor (LF) Inhibitors Using Genetic Algorithms, Pareto Scoring, and Structural Biology. *J. Chem. Inf. Model.* **2012,** 52, 1886-1897.
114. Gozalbes, R.; Barbosa, F.; Nicolai, E.; Horvath, D.; Froloff, N. Development and Validation of a Pharmacophore-Based QSAR Model for the Prediction of CNS Activity. *ChemMedChem* **2009,** 4, 204-209.
115. Mustata, G.; Follis, A. V.; Hammoudeh, D. I.; Metallo, S. J.; Wang, H. B.; Prochownik, E. V.; Lazo, J. S.; Bahar, I. Discovery of Novel Myc-Max Heterodimer Disruptors with a Three-Dimensional Pharmacophore Model. *J. Med. Chem.* **2009,** 52, 1247-1250.
116. Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996,** 36, 563-571.
117. Richmond, N. J.; Abrams, C. A.; Wolohan, P. R. N.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput-Aided. Mol. Des.* **2006,** 20, 567-587.
118. Molecular Operating Environment. http://www.chemcomp.com/.
119. Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput-Aided. Mol. Des.* **2006,** 20, 647-671.
120. Shepphird, J. K.; Clark, R. D. A marriage made in torsional space: using GALAHAD models to drive pharmacophore multiplet searches. *J. Comput-Aided. Mol. Des.* **2006,** 20, 763-771.

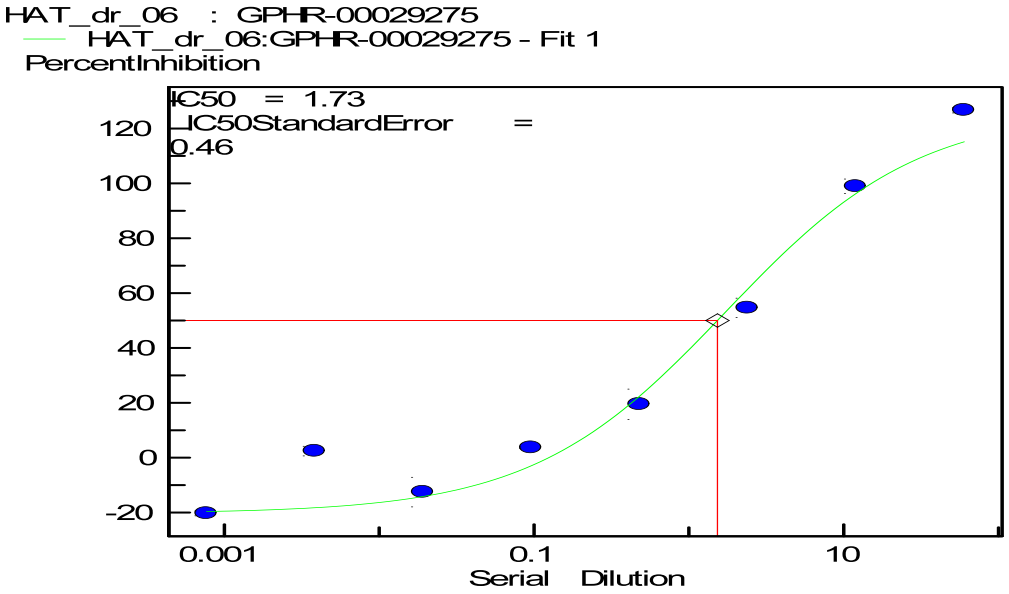
121. Richmond, N. J.; Willett, P.; Clark, R. D. Alignment of three-dimensional molecules using an image recognition algorithm. *J. Mol. Graph. Model.* **2004,** 23, 199-209.
122. Cottrell, S. J.; Gillet, V. J.; Taylor, R.; Wilton, D. J. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *J. Comput-Aided. Mol. Des.* **2004,** 18, 665-682.
123. Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998,** 38, 379-386.
124. Clark, R. D.; Abrahamian, E. Using a staged multi-objective optimization approach to find selective pharmacophore models. *J. Comput-Aided. Mol. Des.* **2009,** 23, 765-771.
125. Pearlman, D. A.; Charifson, P. S. Improved scoring of ligand-protein interactions using OWFEG free energy grids. *J. Med. Chem.* **2001,** 44, 502-511.
126. Vapnik, V. N. *The nature of statistical learning theory*. Springer: New York, 1995; p xv, 188 p.
127. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995,** 20, 273-297.
128. Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000,** 18, 464-477.
129. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010,** 50, 742-754.
130. Gramatica, P. Principles of QSAR models validation: internal and external. *Qsar. Comb. Sci.* **2007,** 26, 694-701.
131. Parker, C. N.; Bajorath, J. Towards unified compound screening strategies: A critical evaluation of error sources in experimental and virtual high-throughput screening. *Qsar. Comb. Sci.* **2006,** 25, 1153-1161.
132. Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; Euler, T. In *Yale: Rapid prototyping for complex data mining tasks*, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006; ACM: 2006; pp 935-940.
133. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002,** 45, 2615-2623.
134. Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput. Aid. Mol. Des.* **2000,** 14, 251-264.
135. Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods .2. Applications. *J. Comput. Chem.* **1989,** 10, 221-264.
136. Stewart, J. J. P. Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements. *J. Mol. Model.* **2004,** 10, 155-164.
137. Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods .3. Extension of Pm3 to Be, Mg, Zn, Ga, Ge, as, Se, Cd, in, Sn, Sb, Te, Hg, Tl, Pb, and Bi. *J. Comput. Chem.* **1991,** 12, 320-341.

138. Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000,** 43, 3714-3717.

APPENDIX ONE:

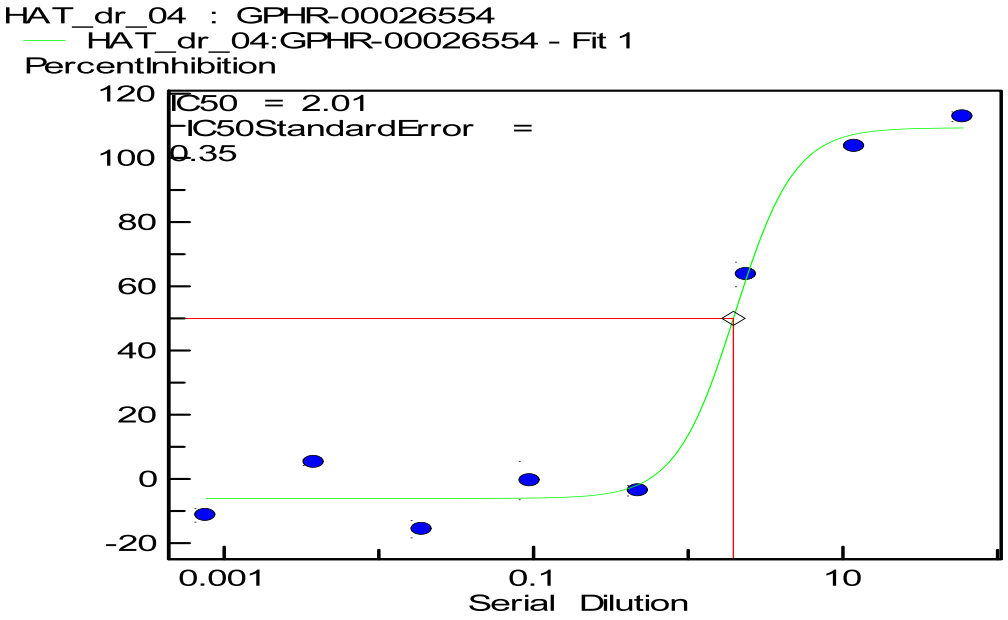
In vitro HTS1 Assay Data for the Fourteen Active, Thiourea-based Rtt109 Inhibitors from the Final 3D-QSAR Compound Training Set Shown in Table 5.1 (Hill: Hill slope which describes the steepness of the dose-response curve)

No. 5.1



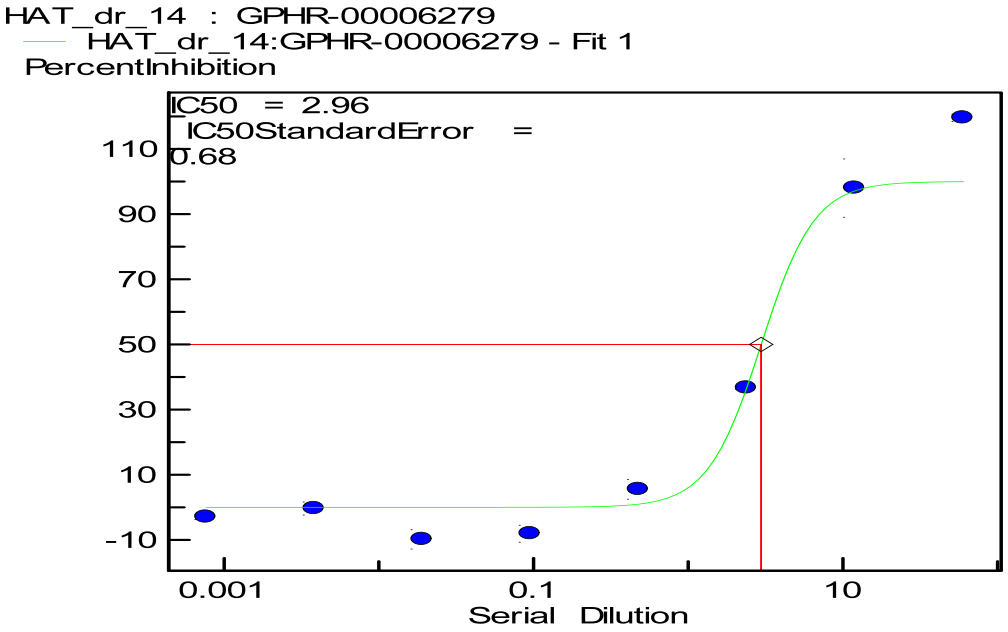
Compound ID	GPHR-00029275
R^2	0.97
Hill	0.69
IC_{50}	1.73 μM
n	2
SE	0.46 μM

No. 5.2



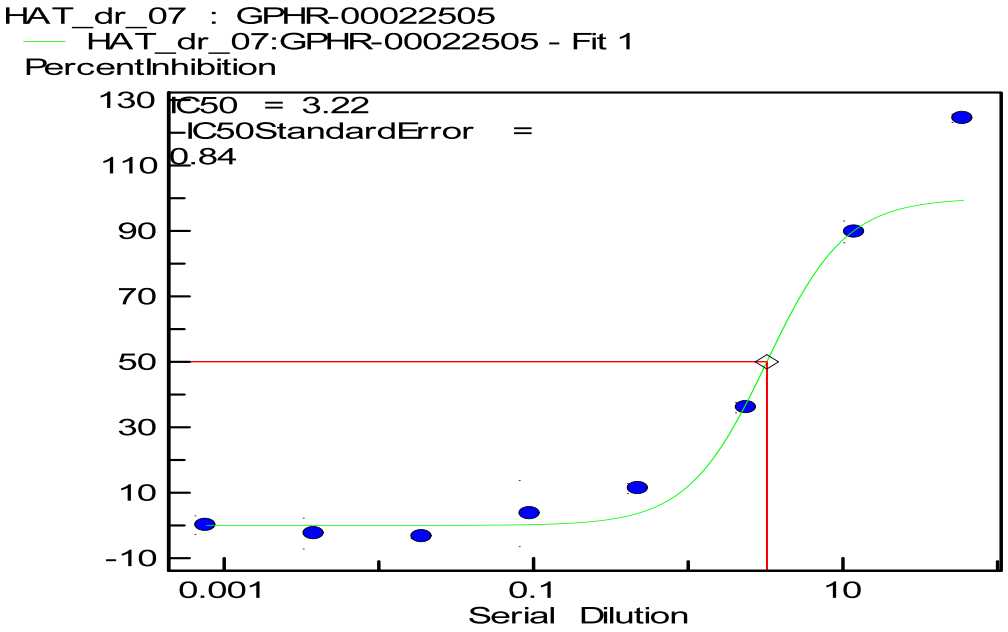
Compound ID	GPHR-00026554
R^2	0.98
Hill	2.26
IC_{50}	2.01 μM
n	2
SE	0.35 μM

No.5.3



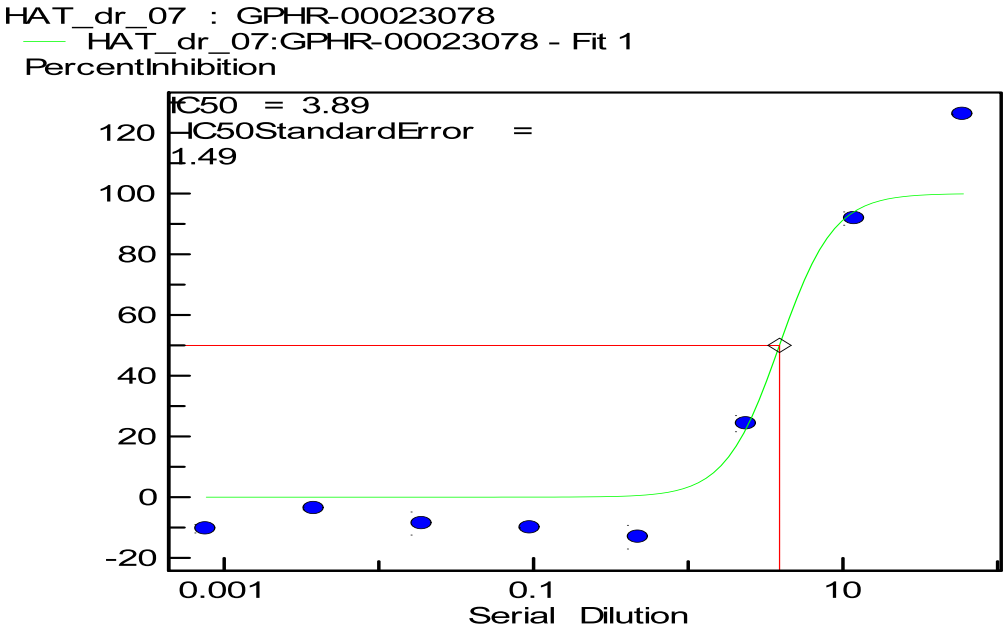
Compound ID	GPHR-00006279
R^2	0.98
Hill	2.53
IC_{50}	2.96 μM
n	2
SE	0.68 μM

No. 5.4



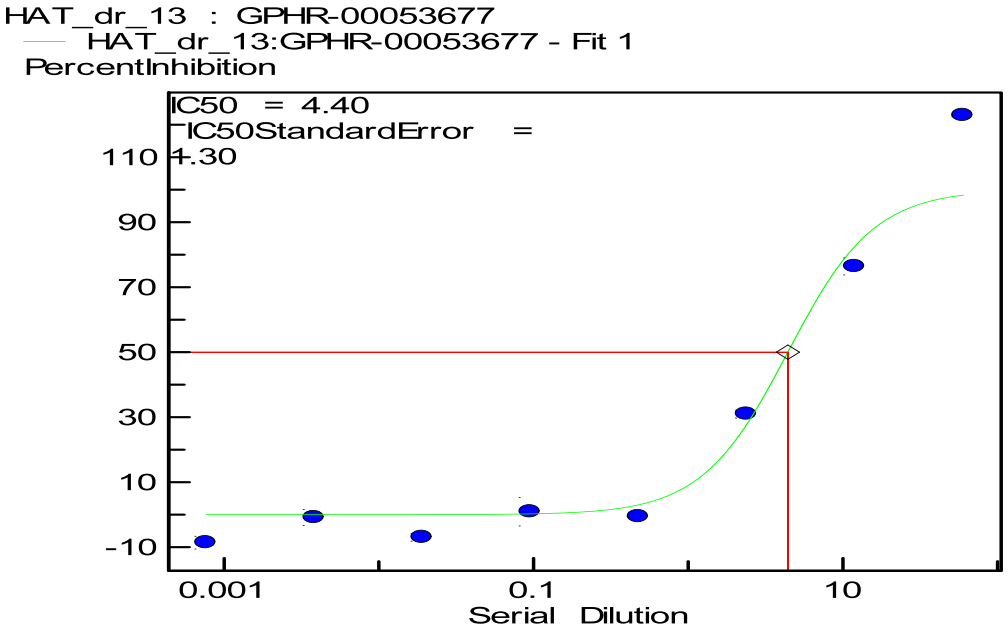
Compound ID	GPHR-00022505
R^2	0.97
Hill	1.70
IC_{50}	3.22 μM
n	2
SE	0.84 μM

No. 5.6



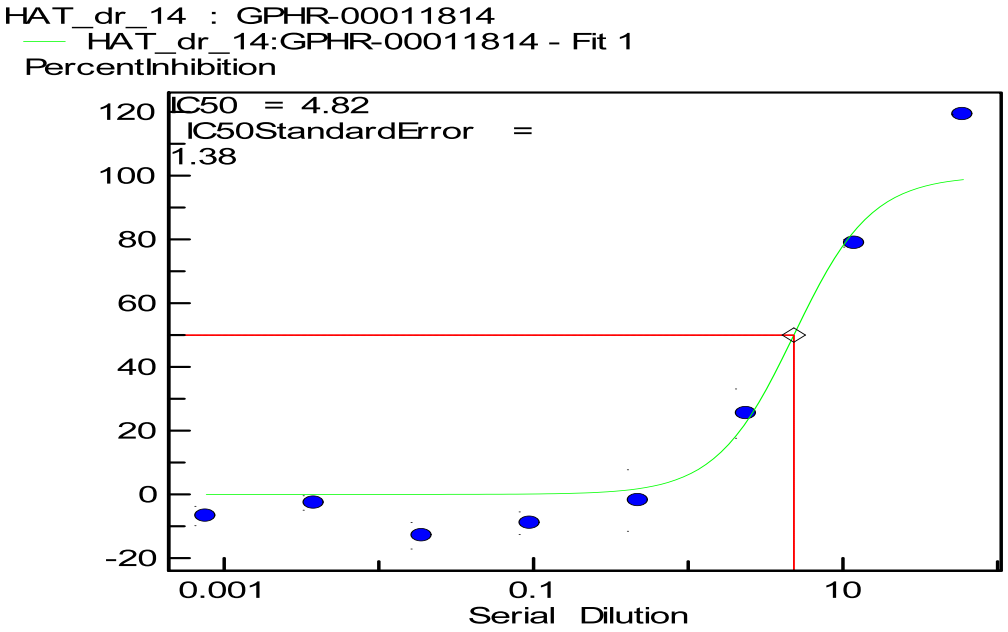
Compound ID	GPHR-00023078
R^2	0.98
Hill	2.48
IC_{50}	3.89 μM
n	2
SE	1.49 μM

No. 5.7



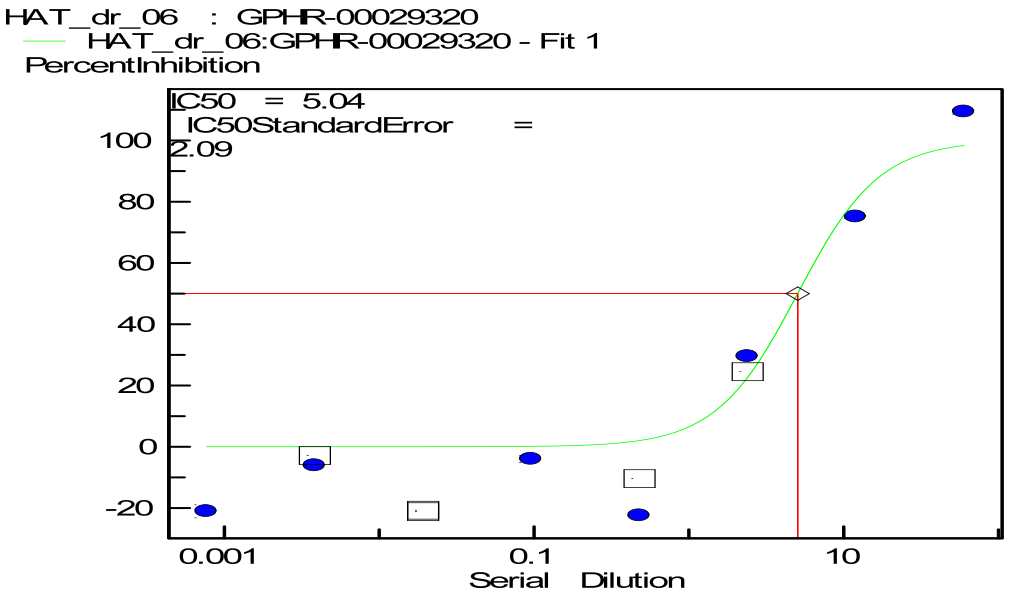
Compound ID	GPHR-00053677
R^2	0.97
Hill	1.56
IC_{50}	4.40 μM
n	2
SE	1.30 μM

No. 5.8



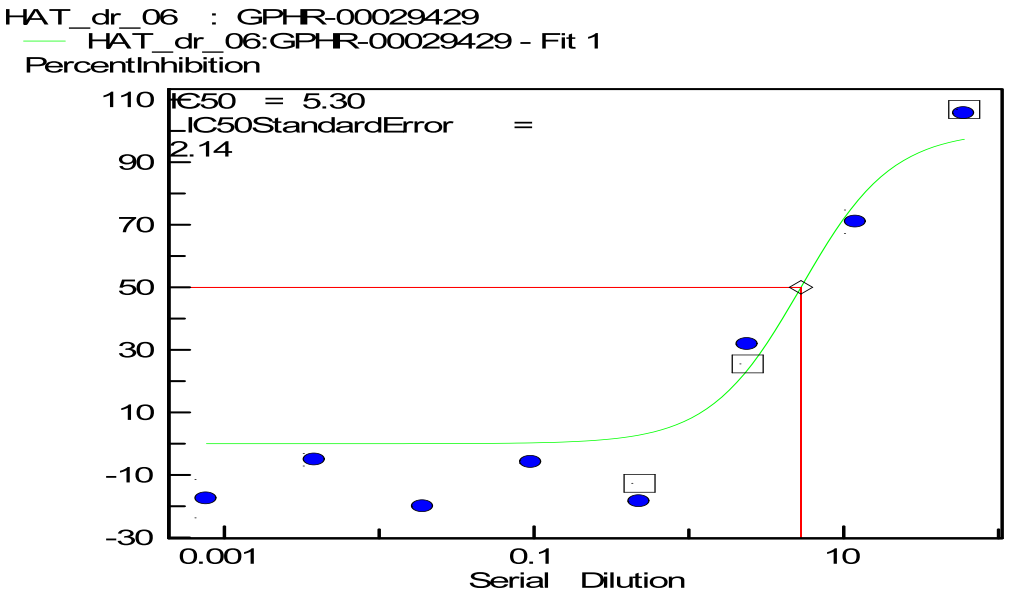
Compound ID	GPHR-00011814
R^2	0.98
Hill	1.73
IC_{50}	4.82 μM
n	2
SE	1.38 μM

No. 5.9



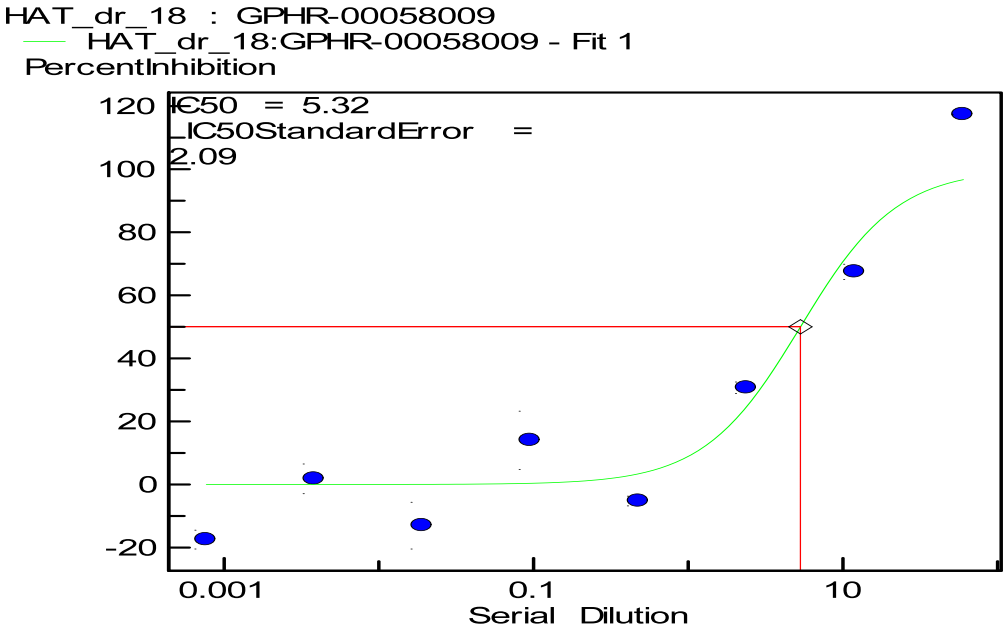
Compound ID	GPHR-00029320
R^2	0.96
Hill	1.65
IC_{50}	5.04 μM
n	2
SE	2.09 μM

No. 5.10



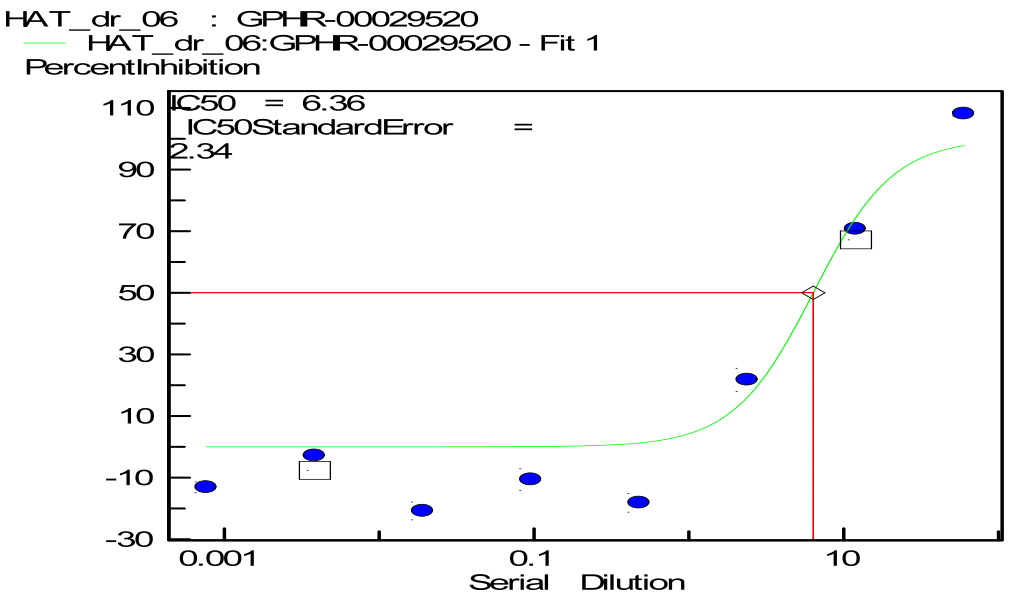
Compound ID	GPHR-00029429
R^2	0.96
Hill	1.48
IC_{50}	5.30 μM
n	2
SE	2.14 μM

No. 5.11



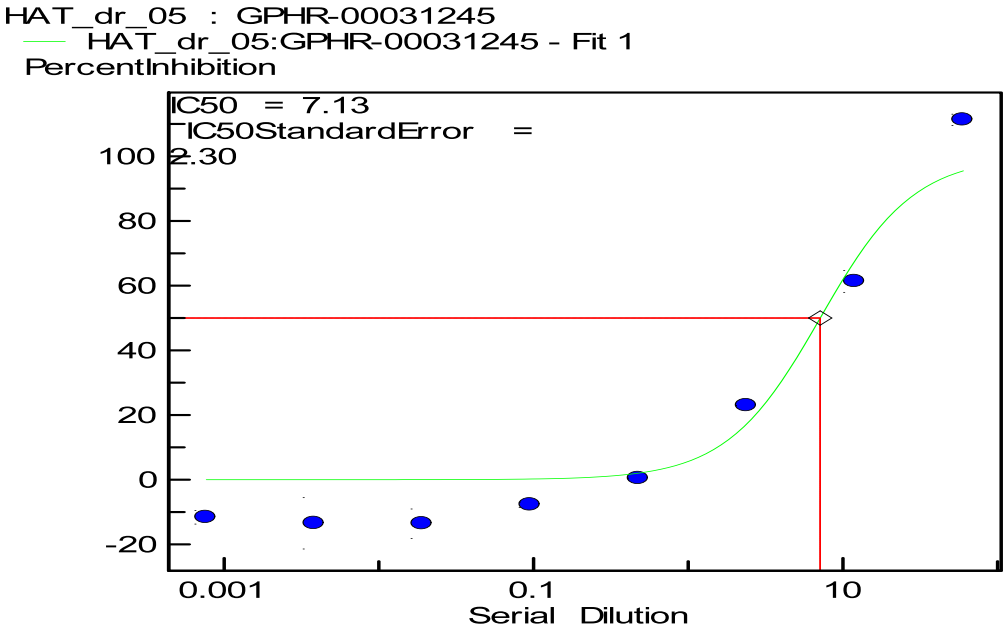
Compound ID	GPHR-00058009
R^2	0.93
Hill	1.39
IC_{50}	5.32 μM
n	2
SE	2.09 μM

No. 5.12



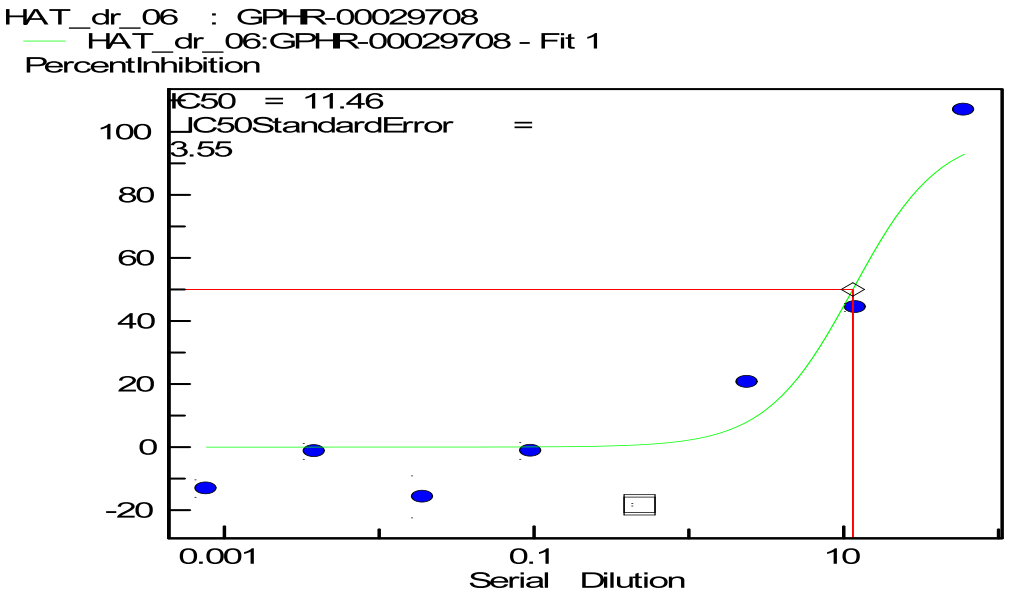
Compound ID	GPHR-00029520
R^2	0.97
Hill	1.69
IC_{50}	6.36 μM
n	2
SE	2.34 μM

No. 5.13



Compound ID	GPHR-00031245
R^2	0.97
Hill	1.43
IC_{50}	7.13 μM
n	2
SE	2.30 μM

No. 5.14

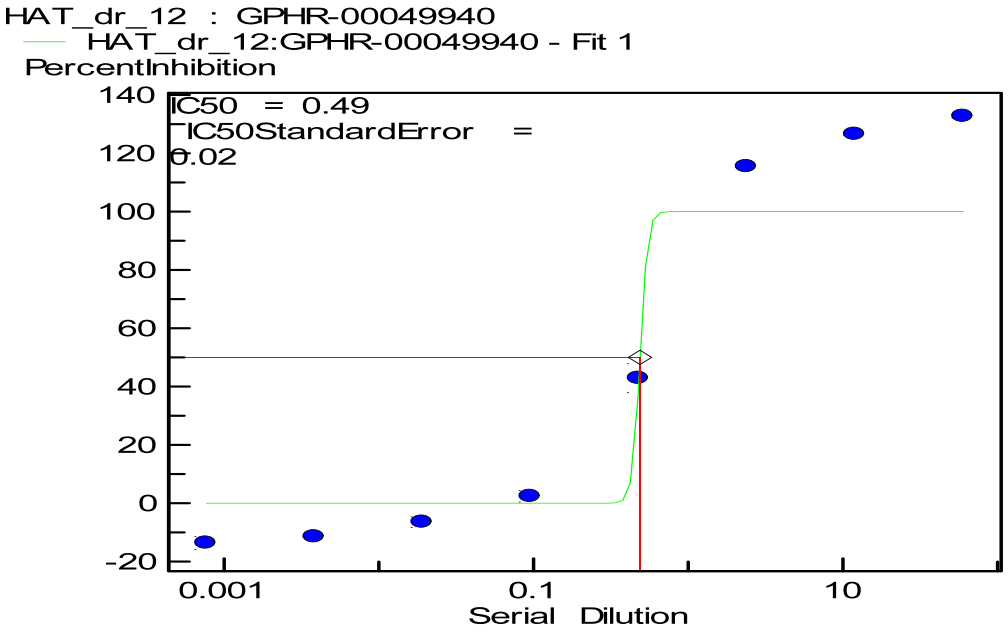


Compound ID	GPHR-00029708
R^2	0.95
Hill	1.55
IC_{50}	11.46 μM
n	2
SE	3.55 μM

APPENDIX TWO:

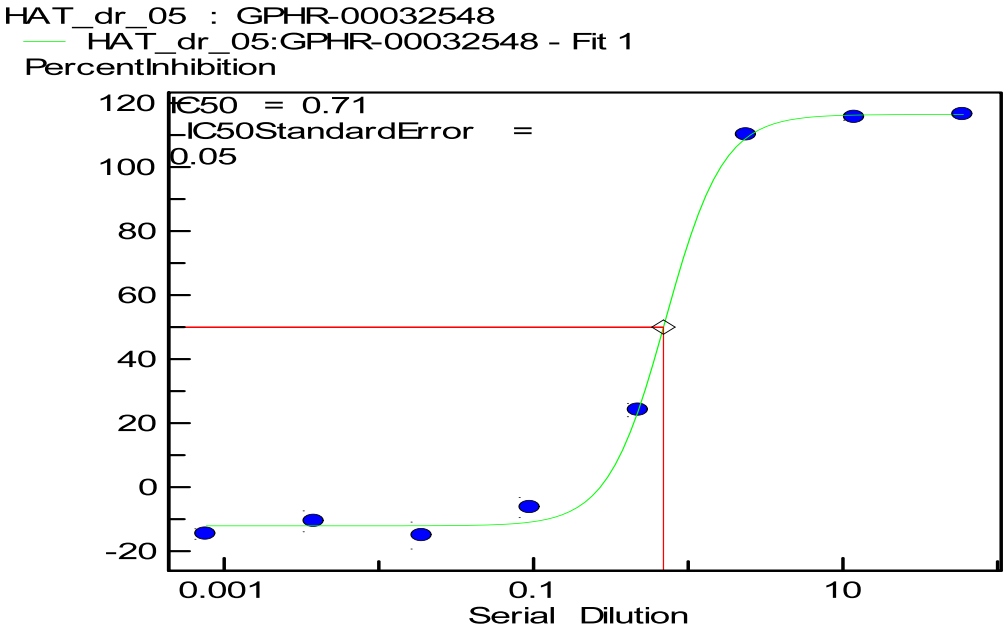
In vitro HTS1 Assay Data for the Five Most Active Compounds Overall from Dataset **ITB,** Used to Generate Pharmacophore Model **H1** (Hill: Hill slope which describes the steepness of the dose-response curve)

No.6.1



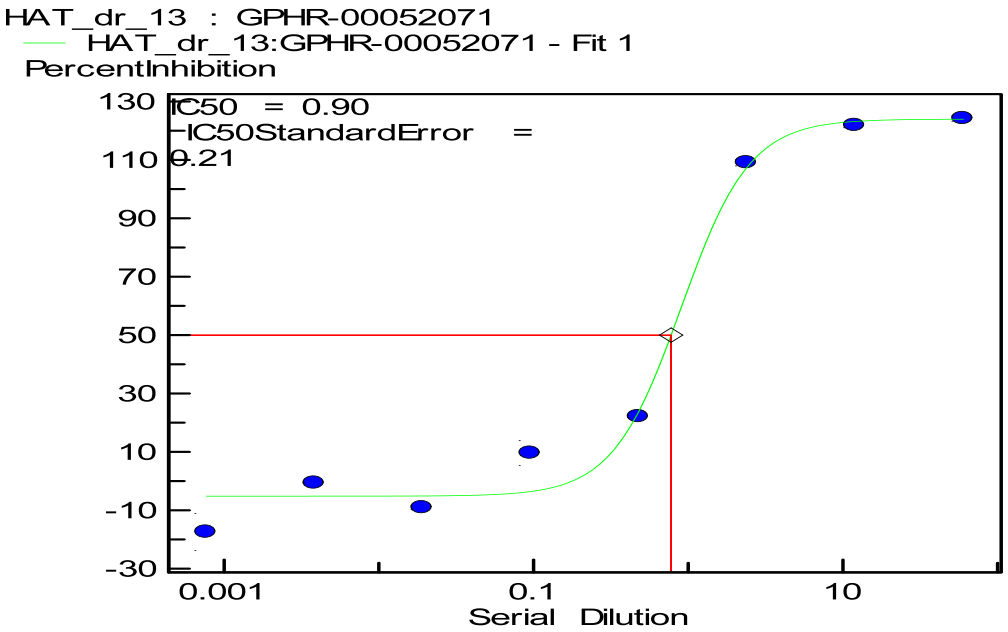
Compound ID	GPHR-00049940
R^2	0.99
Hill	18.05
IC_{50}	0.49 μM
n	2
SE	0.02 μM

No. 6.2



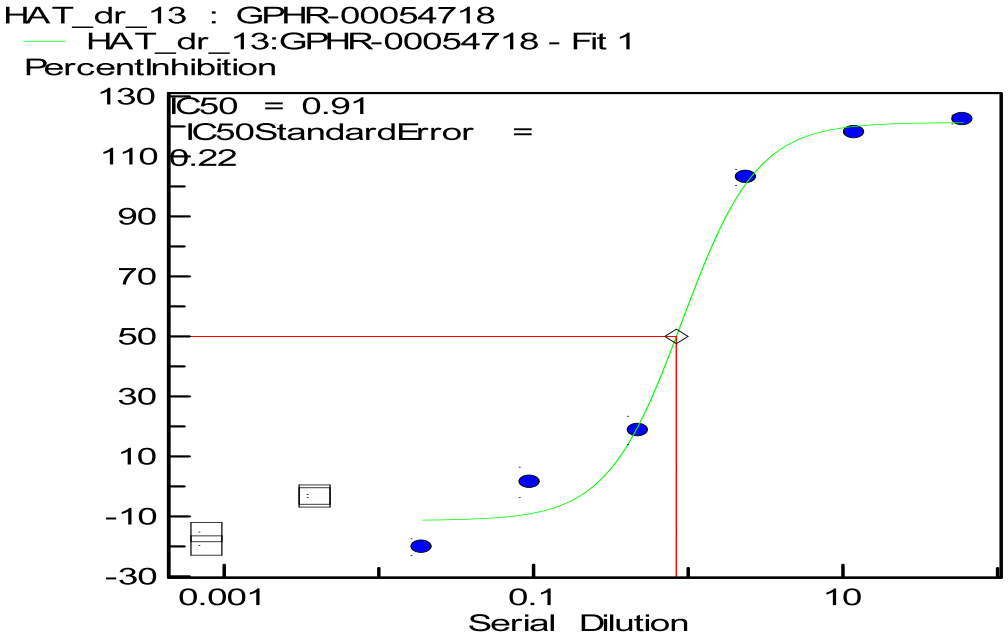
Compound ID	GPHR-00032548
R^2	1.00
Hill	2.33
IC_{50}	0.71 μM
n	2
SE	0.05 μM

No. 6.3



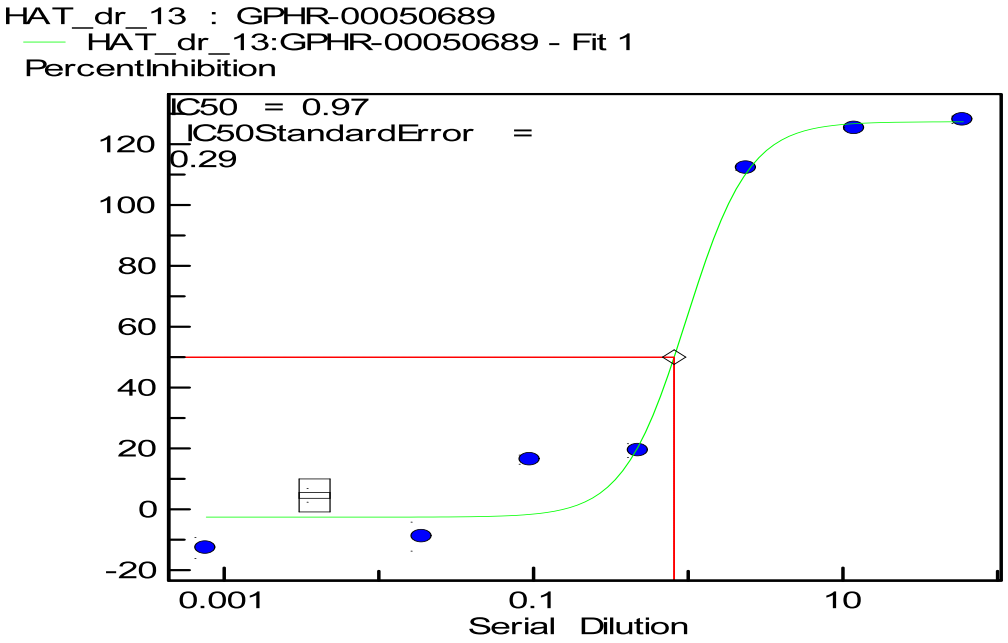
Compound ID	GPHR-00052071
R^2	0.99
Hill	1.96
IC_{50}	0.90 μM
n	2
SE	0.21 μM

No. 6.4



Compound ID	GPHR-00054718
R^2	0.99
Hill	1.79
IC_{50}	0.91 μM
n	2
SE	0.22 μM

No. 6.5



Compound ID	GPHR-00050689
R^2	0.98
Hill	2.13
IC_{50}	0.97 μM
n	2
SE	0.29 μM