

# 國立交通大學

生物資訊所

碩士論文

LigSeeSVM: 結合 Support Vector Machines 與資料融合在活性配體為基之藥物篩選及 GPCR 與 GABA<sub>A</sub> 之實際應用

LigSeeSVM: Support Vector Machines and Data fusion for Ligand-based Compound Screening and Applications to GPCR and GABA<sub>A</sub>

研究生：林柏村

指導教授：楊進木 博士

中華民國九十四年六月

# LigSeeSVM：結合 Support Vector Machines 與資料融合在活性配體為基之藥物篩選及 GPCR 與 GABA<sub>A</sub> 之實際應用

學生：林柏村

指導老師：楊進木

國立交通大學生物資訊所碩士班

## 摘要

以配體為基(ligand-based)的藥物設計乃是因為受體(target protein)結構尚未被解出亦或無法得知，因而以活性配體的結構作分析歸納。本論文研究，主要是以 2 組不同的描述子(descriptor)來描述化合物特性：(1)atom-pair 描述子(AP descriptor) 825 個、(2)藥物熱動力學描述子(thermodynamic descriptor)6 個及 Accelrys Cerius<sup>2</sup> 預設的描述子(Cerius<sup>2</sup> default descriptor) 13 個。應用這 2 組不同的描述子，以 LibSVM 為篩選工具，分別產生 2 組不同的結果，將 SVM 輸出的結果，轉換成 Z-score，再將結果依 Z-score 排序。最後，更將 2 組結果，用 rank 資料融合(rank combination)的方法，得到第 3 組結果。本研究以 TK (thymidine kinase) 活性配體、ER 抑制劑(estrogen receptor antagonist)、ER 促進劑(estrogen receptor agonist) 各 10 個、GPCR 及 GABA<sub>A</sub> 活性配體共 100 個，再加上從化合物資料庫 ACD 中隨機挑選出 990 個及化合物資料庫 CMC 中隨機挑選出 7300 個化合物，做為測試組，以 SVM 預測已知活性配體在所有化合物中的位置，藉此觀察 SVM 在篩選化合物資料庫上的表現。在以 990 個 ACD 化合物為化合物資料庫的測試組裡，比較 SVM 與其他方法(Surflex-Sim, Ajay N. Jain, 2004)的表現，SVM 在以 rank 資料融合方法所得到的結果為最好，在 true hit% 達 100% 時，偽陽性的比例(false positive rate) 分別為 0.3% (TK)、0.6% (ER antagonists) 及 0% (ER agonists)。以 ROC 曲線圖而言，在 GPCR 及 GABA<sub>A</sub> 活性配體測試組的表現，SVM 也確實在虛擬藥物篩選上表現較好。在以 7300 個 CMC 化合物為化合物資料庫的測試組裡，我們觀察到擁有較高的 Z-score 且排名在前面的化合物中，大部分與已知活性配體具有極其相似的結構。有一部分化合物擁有高 Z-score，但其結構與已知活性配體不相似，很有可能是新的先導化合物(novel lead compounds)。綜合 SVM 在上述不同測試組的結果，我們可以確定 SVM 是適合用於虛擬藥物篩選上並且其表現優於其他同樣以活性配體為基的方法。

LigSeeSVM: Support Vector Machines and Data fusion for Ligand-based Compound  
Screening and Applications to GPCR and GABA<sub>A</sub>

Student: Po-Tsun Lin

Advisor: Jinn-Moon Yang

Institute of Bioinformatics  
National Chiao Tung University

ABSTRACT

A major benefit of ligand-based drug screening approaches is that they can perform screening even though the drug targets whose three-dimensional structure is not known enough to permit structure-based virtual screening. In this thesis, we have developed a **Ligand-based Screening** tool using Support Vector Machines and data fusion method, termed LigSeeSVM. We combine structure descriptors (825 atom pair descriptors) and physicochemical descriptors (Accelrys Cerius<sup>2</sup> six thermodynamic and 13 default descriptors) to characterize compounds' features. Next, we used SVM to generate SVM-AP model based on 825 AP descriptors and SVM-PC model based on 19 physicochemical descriptors. The predicted scores of both SVM-AP and SVM-PC models are normalized by transferring the scores to Z-scores. We fused SVM-AP and SVM-PC predicted results using rank combination to create LigSeeSVM predicted model, respectively. In this study, we used 10 thymidine kinase substrates, 11 estrogen receptor antagonists, 10 estrogen receptor agonists, 100 GPCR and GABA<sub>A</sub> ligands, combined with 990 randomly chosen compounds from the ACD or 7300 randomly chosen compounds from the CMC as screening sets. Using these screening sets to verified the utility of LigSeeSVM on virtual screening. When the true hit rate was 100%, the false positive rates were 0.3% for TK, 0.6% for ER antagonists, and 0% for ER agonists. The ROC curves of GPCR and GABA<sub>A</sub> screening sets also shown that the performance of the LigSeeSVM is better than other ligand-based virtual screening approaches on these data sets. The results of the LigSeeSVM using 7300 CMC randomly chosen compounds as compound database shown that the majority of compounds with high Z-score also have structures similar to the known ligands, some compounds with high Z-score but have different structures compared with the known ligands, and these compounds have more possibility to become novel, potential lead compounds. Our results suggest that LigSeeSVM is practically applicable for ligand-based virtual screening and offers competitive performance to other ligand-based virtual screening approaches on these data sets.

## **Acknowledgements**

The most appreciation is for my advisor Dr. Jinn-Moon Yang. Because of his advice and introduction, I could have a start to learn about bioinformatics and finally finish the thesis. I am very grateful for his knowledge and suggestions that help me to correct my mistakes. Without his guides and encouragement, it is impossible to finish the thesis during these two years. Thanks for all members in my laboratory, this is a perfect team that we could discuss each other to solve someone's problem. Finally I would like to thank my parents and family for supporting me through these two years.



## CONTENTS

Abstract (in Chinese).....	I
Abstract.....	II
Acknowledgements .....	III
Contents .....	IV
List of Tables .....	VI
List of Figures.....	VIII
Chapter 1. Introduction.....	01
1.1 Motivations and Purposes.....	01
1.2 Thesis Overview .....	03
Chapter 2. Materials and Methods.....	05
2.1 Preparation of ACD Screening Databases.....	05
2.2 Preparation of CMC Screening Databases .....	06
2.3 Feature Extraction.....	07
2.4 Build LigSeeSVM prediction Model.....	12
Chapter 3. Evaluation LigSeeSVM on ACD Database .....	14
3.1 Thymidine Kinase.....	14
3.2 Estrogen Receptor.....	16
3.3 GPCR and GABA <sub>A</sub> Receptor .....	17

Chapter 4. Evaluation LigSeeSVM on CMC Database.....	19
4.1 Thymidine Kinase.....	19
4.2 Estrogen Receptor.....	20
4.3 GPCR and GABA <sub>A</sub> Receptor .....	21
Chapter 5. Conclusions.....	22
5.1 Summary.....	22
5.2 Major Contributions and Future Perspectives .....	22
References .....	60



## List of Tables

<b>Table 1.</b> 100 positive compounds of GPCR and GABA <sub>A</sub> studied in this thesis [3].....	<b>24</b>
<b>Table 2.</b> Ten atom types used for atom pair descriptors [11] .....	<b>27</b>
<b>Table 3.</b> Six thermodynamic descriptors available in Accelrys Cerius <sup>2</sup> QSAR+ module [13] .....	<b>28</b>
<b>Table 4.</b> 13 default descriptors available in Accelrys Cerius <sup>2</sup> QSAR+ module [13] .....	<b>29</b>
<b>Table 5.</b> Accuracies of three kinds SVM models for 8 TK test substrates .....	<b>30</b>
<b>Table 6.</b> Comparison our three SVM models with Surflex-Sim on false positive rates for thymidine kinase substrates and ACD database at true positive rates 80% and 100% .....	<b>31</b>
<b>Table 7.</b> Accuracies of three kinds SVM models for 9 ER test antagonists.....	<b>32</b>
<b>Table 8.</b> Comparison our three SVM models with Surflex-Sim on false positive rates for estrogen receptor antagonists and ACD database at true positive rates 80% and 100% ...	<b>33</b>
<b>Table 9.</b> Accuracies of three kinds SVM models for 9 ER test antagonists .....	<b>34</b>
<b>Table 10.</b> Top ranked 10 compounds of our three SVM models on screening ER agonists from the ACD chemical database.....	<b>35</b>
<b>Table 11.</b> The remaining 5 TK known ligands in which they ranked in a screen including 7100 randomly chosen compounds from CMC.....	<b>36</b>
<b>Table 12.</b> Top ranked 10 compounds of the LigSeeSVM on screening TK ligands from the CMC chemical database .....	<b>37</b>
<b>Table 13.</b> The remaining 6 ER antagonists in which they ranked in a screen including 7100 randomly chosen compounds from CMC.....	<b>38</b>
<b>Table 14.</b> Top ranked 10 compounds of the LigSeeSVM on screening ER antagonists from the CMC chemical database .....	<b>39</b>
<b>Table 15.</b> The remaining 5 ER agonists in which they ranked in a screen including 7100 randomly chosen compounds from CMC.....	<b>40</b>

<b>Table 16.</b> Top ranked 10 compounds of the LigSeeSVM on screening ER agonists from the CMC chemical database .....	<b>41</b>
<b>Table 17.</b> Top ranked 10 compounds of the LigSeeSVM on screening serotonin receptor ligands from the CMC chemical database .....	<b>42</b>
<b>Table 18.</b> Top ranked 10 compounds of the LigSeeSVM on screening muscarinic receptor ligands from the CMC chemical database .....	<b>43</b>
<b>Table 19.</b> Top ranked 10 compounds of the LigSeeSVM on screening histamine receptor ligands from the CMC chemical database .....	<b>44</b>
<b>Table 20.</b> Top ranked 10 compounds of the LigSeeSVM on screening GABA <sub>A</sub> receptor ligands from the CMC chemical database .....	<b>45</b>





## List of Figures

- Figure 1.** Overview of our methods for ligands-based compound screening .....46
- Figure 2.** The positive compounds in training sets of GPCR and GABA<sub>A</sub> .....47
- Figure 3.** The general definition of atom pair descriptors .....48
- Figure 4.** 10 thymidine kinase inhibitors were studied for ligand-based screening .....49
- Figure 5.** (A) The UPGMA rooted tree of 10 TK inhibitors. (B) The UPGMA rooted tree of 11 ER antagonists. (C) The UPGMA rooted tree of 10 ER agonists..... 50
- Figure 6.** The (A) true hits and (B) GH scores of our three SVM models (SVM-AP, SVM-PC, and LigSeeSVM) in screening a TK set with eight positive substrates and 950 negative compounds..... 51
- Figure 7.** 11 estrogen receptor antagonists were studied for ligand-based screening.....52
- Figure 8.** 10 estrogen receptor agonists were studied for ligand-based screening .....53
- Figure 9.** The (A) true hits and (B) GH scores of our three SVM models (SVM-AP, SVM-PC, and LigSeeSVM) in screening an ER set with nine positive antagonists and 950 negative compounds..... 54
- Figure 10.** The (A) true hits and (B) GH scores of our three SVM models (SVM-AP, SVM-PC, and LigSeeSVM) in screening an ER set with seven positive agonists and 950 negative compounds ..... 55
- Figure 11.** (A) ROC curve of our three SVM models for 43 muscarinic receptor ligands. (B) ROC curve of Surflex-Sim and Tanimoto [3] methods for 43 muscarinic receptor ligands. The LigSeeSVM model performs better than Surflex-Sim ..... 56
- Figure 12.** (A) ROC curve of our three SVM models for 47 histamine receptor ligands. (B) ROC curve of Surflex-Sim and Tanimoto [3] methods for 47 histamine receptor ligands. The LigSeeSVM model performs better than Surflex-Sim ..... 57
- Figure 13.** (A) ROC curve of our three SVM models for 15 GABA<sub>A</sub> receptor ligands. (B) ROC curve of Surflex-Sim and Tanimoto [3] methods for 15 GABA<sub>A</sub> receptor ligands. The

LigSeeSVM model performs better than Surfex-Sim ..... **58**

**Figure 14.** (A) ROC curve of our three SVM models for 29 serotonin receptor ligands. (B)

ROC curve of Surfex-Sim and Tanimoto [3] methods for 29 serotonin receptor ligands ..

..... **59**



## Chapter 1

### Introduction

#### 1.1 Motivations and Purposes

The pharmaceutical industry is under ever-increasing pressure to increase its success rate in bringing drugs to the market. It is estimated that on average it can take 14 years to bring a compound from hit identification through to an approved drug [1], and the costs associated with this process are enormous. In recent years, because the development of computer hardware and software, computational screening of compound databases has become increasingly popular in pharmaceutical research. It will save much time and cost to find novel, potential inhibitors for diseases with aids of computer. The computational approaches used for virtual screening can be classified into two categories: structure-based screening (often referred to as docking) and screening using active compounds as templates (ligand-based virtual screening). For ligand-based methods, the strategy is to use information provided by a compound or set of compounds that are known to bind to the desired target and to use this to identify other compounds in the corporate database or external databases with similar properties [1].

Currently, the applications of structure-based virtual screening approaches relying on a detailed three-dimensional model of the receptor binding pocket [2], but some drug targets for small molecule therapeutics are proteins whose three-dimensional structures are not known to permit structure-based virtual screening [3]. For example, most membrane spanning G-protein-coupled receptors (GPCRs) or ion channels are 3D structures unavailable. Besides, GPCRs or ion channels were the targets for nine of the

top20 selling prescription drugs worldwide in the year 2000 [3, 4]. Therefore, we followed an entirely ligand-based approach to GPCRs and GABA<sub>A</sub> receptors.

In the study of pharmacological properties of drugs and other chemical agents, a variety of molecular descriptors have been developed and routinely used for describing physicochemical and structural properties of chemical agents [5-9]. These descriptors were initially developed for the construction of quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) of structurally related compounds [5, 10]. There are many different approaches to generate descriptors. These include both 2D and 3D methods. Most of the 2D methods are based upon graph theoretic indices (structural indices). Although these structural indices represent different aspects of molecular structures, their physicochemical meaning is unclear. Besides, 2D methods cannot distinguish stereoisomers [11]. A major benefit of 2D methods is that the former neither requires conformational search nor structural alignment. Accordingly, 2D methods are easily automated and adapted to task of database searching, or virtual screening [12]. The major molecular descriptors used in this work are derived from 2D molecular topology (825 different atom pair descriptors) [11]; however, for the disadvantages of the 2D methods, we used secondary kind of descriptors: Accelrys Cerius<sup>2</sup> QSAR module 6 thermodynamic and 13 default descriptors [13].

Support vector machines (SVMs) have been applied to a wide range of pharmacological and biomedical problems including drug-likeness, drug blood-brain barrier penetration prediction, drug-receptor binding, and drug metabolism [5, 14-18]. In this study, support vector machines (SVMs) were used for virtual screening because they are known to be a powerful technique. The theory of SVMs has been extensively

described in many literatures [5]. Thus only a brief description is given here. SVMs are supervised learning algorithms proposed by Vapnick (1995). Data examples labeled as positive or negative are projected into high-dimensional feature space using a kernel, and the hyper-plane in the feature space is optimized to maximize the margin between the positive and negative examples [19].

The results of this study suggests that our approach, which combines SVMs and virtual screening, can be explored as a general drug discovery tool and applied to a large variety of available datasets of biologically active compounds.

## 1.2 Thesis Overview

We have used LibSVM 2.71 (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>), developed by Lin et al. [20], on virtual database screening. In chapter 2, we used 10 thymidine kinase substrates, 11 estrogen receptor antagonists, 10 estrogen agonist, 100 GPCR and GABA<sub>A</sub> ligands and combined with 990 randomly chosen compounds from ACD or 7300 randomly chosen compounds from CMC to form fourteen screening sets. After preparing screening sets, we extracted two kinds of features to describe physicochemical and structural properties of compounds. Next, we used LibSVM 2.71 with different screening sets to evaluate the performance on virtual screening.

In chapter 3, we evaluated the screening performance of LibSVM with screening sets (TK, ER antagonists, ER agonists, GPCR, and GABA<sub>A</sub> ligands) combined with 990 ACD randomly chosen compounds by the true hit, hit rate, goodness-of-hit (GH) and ROC curves. We used three combinations of atom pair descriptors only, physicochemical descriptors only, and rank combination to compare screening utilities with other ligand-based virtual screening approach.

The results showed that the rank combination was the most reliable.

In chapter 4, we applied the LibSVM with screening sets (TK, ER antagonists, ER agonists, GPCR, and GABA<sub>A</sub> ligands) combined with 7300 CMC randomly chosen compounds. The results showed that the majority of compounds with high Z-score have structures similar to the known ligands. Some compounds with high Z-score but have different structures compared with the known ligands, and these compounds have more possibility to become novel, potential lead compounds.

Chapter 5 presented some conclusions and future perspectives. Our results suggest that SVM is practically applicable for virtual screening and offer competitive performance to other ligand-based virtual screening approach. Combination has great improved the result of drug screening approach and it could effectively reduce the number of false positives.



## Chapter 2

### Methods and Materials

The SVM method was proposed by Vapnik [21, 22] on the basis of the Structural Risk Minimization Principle [21, 23]. It was initially designed to solve pattern recognition problems [21, 24], but it was later applied to function estimation problems [21, 25]. The estimated function is a linear expansion in terms of functions defined on a certain subset of the data (support vectors), and the final number of coefficients in such an expansion does not depend on the dimensionality of the space of input variables. These two properties make SVM an especially useful technique for dealing with very large data sets in a high-dimensional space [21].

In this study, we used LibSVM 2.71 (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) [20]. Data examples labeled as positive or negative are projected into a high-dimensional feature space and the hyper-plane in the feature space is optimized to maximize the margin between the positive and negative [19]. The general flowchart of the virtual screening using LibSVM is shown in Figure 1.

### 2.1 Preparation of ACD Screening Databases

#### A. Thymidine kinase substrates and estrogen receptor antagonists and agonists

Two screening sets designed for virtual screening against TK and ER $\alpha$  antagonists were proposed by Bissantz et al. [26] in 2000 and retested by Jain in 2003 [27]. A TK library contained 10 known ligands of TK and 990 randomly chosen molecules from the ACD; an ER $\alpha$ -antagonists library contained 11 known antagonists of ER $\alpha$  and 990 randomly chosen molecules from the ACD. The two sets were downloaded from

<http://jainlab.ucsf.edu/Downloads.html> proposed by Jain and the files in the SYBYL mol2 format were converted to the MDL mol format with Corina 3.0 [28] for running Accelrys Cerius<sup>2</sup> QSAR module [13]. Besides, the screening set for ER $\alpha$  agonists included 10 known agonists [29] and the same 990 molecules as the ER $\alpha$ -antagonists library.

## B. GPCR and GABA<sub>A</sub> ligands

Serotonin, muscarinic, histamine, and GABA<sub>A</sub> ligands (shown in Figure 2) [3] combined with 40 randomly chosen molecules from the 990 ACD molecules that were used for SVM training set. A and B are serotonin ligands [30-32]. C (tolterodine), D, and E are muscarinic antagonists [33-35]. Molecules F–H (bromodiphenhydramine, pyrilamine, and azatadine) are H1 receptor antagonists. Molecules I – K are GABA<sub>A</sub> receptor agonists (diazepam, alprazolam, and zopiclone).

To test the models that resulted from these input structures, GPCRDB ([www.gpcr.org](http://www.gpcr.org)) [3, 36] was used to identify 85 GPCR ligands of widely varying chemical structure and 15 GABA<sub>A</sub> agonists, of which nine were variations on the classic benzodiazepine scaffold and six were of varying chemotypes. Table 1 lists the names and annotated target specificity for all 100 molecules [3]. As a negative control for all cases, the 990 ACD randomly chosen molecules described above were used.

### 2.2 Preparation of CMC Screening Databases

The larger screening set was derived from the drug database, Comprehensive Medicinal Chemistry (CMC). Using ISIS, the CMC were first filtered with molecular weights between 200 and 500 to yield about 7300 molecules and then removed small fragments and added hydrogen atoms with Corina 3.0. The structure files of molecules were stored in the MDL mol

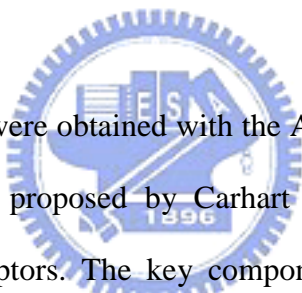


format for running Accelrys Cerius<sup>2</sup> QSAR module and in the SYBYL mol2 format for running AP generator.

A TK library contained 10 known TK substrates and 7300 molecules from the CMC; an ER $\alpha$ -antagonists library contained 11 known antagonists of ER $\alpha$  and 7300 molecules from the CMC; the screening set of ER $\alpha$  agonists contained 10 known agonists and 7300 molecules from the CMC. Besides, the screening set of 85 GPCR and 15 GABA<sub>A</sub> ligands also included the same 7300 CMC molecules.

## 2.3 Feature Extraction

### A. Atom pair descriptors



Atom pair descriptors (AP) were obtained with the AP generator program developed in our laboratory using an approach proposed by Carhart et al. [12, 37]. Figure 3 shows the definition of atom pair descriptors. The key components for defining a set of atom pair descriptors include the definition of atom types and the topological distance bins. An atom pair is a simple type of substructure defined in terms of the atom types and the shortest path separation (or graph distance) between two atoms. The graph distance is defined as the smallest number of atoms along the path connecting two atoms in a molecular structure. The general form of an atom pair is as follows:

$$\text{Atom type } i \text{ --(distance)-- atom type } j$$

where (distance) is the graph distance between atoms  $i$  and  $j$  in the case of a 2D atom pair description. (The distance can also be defined as the physical distance between atom  $i$  and  $j$  in the case of a 3D atom pair description.)

In this study, SYBYL atom types (mol2 format) were utilized as the starting point. In principle, all SYBYL atom types can be used in the generation of atom pair descriptors. In order to reduce the number of atom pair descriptors and improve the accuracy, we have clustered 23 atom types into 10 atom types (Table 2). The total number of pairwise combinations of all 10 atom types is 55. Furthermore, 15 distance bins were defined in the interval of graph distance rang from zero (i.e., zero atoms separating an atom pair) to 14. Thus, a total of 825 (55 × 15) atom pair descriptors were generated for each molecular structure [11].

## B. Accelrys Cerius<sup>2</sup> QSAR module default and thermodynamic descriptors

The physicochemical descriptors of a compound are generated by using Accelrys Cerius<sup>2</sup> QSAR module. QSAR module provides a wide variety of descriptors, in this study, we used two functional families of descriptors, six descriptors in the thermodynamic family (Table 3) and 13 descriptors in the default family (Table 4) [13]. Each descriptor is briefly described as follows:

**Sum of atomic polarizabilities (Apol):** The sum of atomic polarizabilities (Apol) descriptor computes the sum of the atomic polarizabilities. The polarizabilities are caculated from the A coefficients used for molecular mechanics calculations:

$$P_a = \sum_i A_i$$

For more information, see Marsali and Gasteiger (1980); Hopfinger (1973).

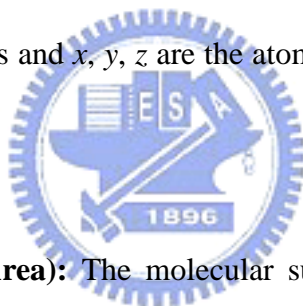
**Dipole moment (Dipole):** The dipole moment descriptor is a 3D electronic descriptor that indicates the strength and orientation behavior of a molecule in an electrostatic field. Both the

magnitude and the components (X, Y, Z) of the dipole moment are calculated. It is estimated by utilizing partial atomic charges and atomic coordinates. Partial atomic charges are computed using the charge setup option in the QSAR control panel offering CHARMM charging rules, Gasteiger, CNDO2, and Del Re methods. The descriptor uses Debyes units. Dipole properties have been correlated to longrange ligand-receptor recognition and subsequent binding.

**Radius of gyration (RadOfGyration):** The radius of gyration is calculated using the following equation:

$$Rog = \sqrt{\left( \sum \frac{(x_i^2 + y_i^2 + z_i^2)}{N} \right)}$$

where  $N$  is the number of atoms and  $x, y, z$  are the atomic coordinates relative to the center of mass.



**Molecular surface area (Area):** The molecular surface area descriptor is a 3D spatial descriptor that describes the van der Waals area of a molecule. The molecular surface area determines the extent to which a molecule exposes itself to the external environment. This descriptor is related to binding, transport, and solubility.

**Molecular weight (MW):** Molecular weight.

**Molecular Volume (Vm):** A 3D spatial descriptor that defines the molecular volume inside the contact surface. The molecular volume is calculated as a function of conformation. Molecular volume is related to binding and transport.

**Density (Density):** A 3D spatial descriptor that is defined as the ratio of molecular weight

to molecular volume. It has the units of  $\text{g ml}^{-1}$ . The density reflects the types of atoms and how tightly they are packed in a molecule. Density can be related to transport and melt behavior.

**Principal moment of inertia (PMI):** Calculates the principal moments of inertia about the principal axes of a molecule according to the following rules:

- ◆ The moments of inertia are computed for a series of straight lines through the center of mass. The moments of inertia are given by:

$$I = \sum_i m_i d_i^2$$

- ◆ Distances are established along each line proportional to the reciprocal of the square root of  $I$  on either side of the center of mass. The locus of these distances forms an ellipsoidal surface. The principal moments are associated with the principal axes of the ellipsoid.
- ◆ If all three moments are equal, the molecule is considered to be a symmetrical top. If no moments are equal, the molecule is considered to be an unsymmetrical top.

For more information about this descriptor, see Hill (1960).

**Number of rotatable bonds (Rotlbonds):** Counts the number of bonds in the current molecule having rotations that are considered to be meaningful for molecular mechanics. All terminal H atoms are ignored (for example, methyl groups are not considered rotatable).

**Hbond acceptor:** Number of hydrogen-bond acceptors.

**Hbond donor:** Number of hydrogen-bond donors.

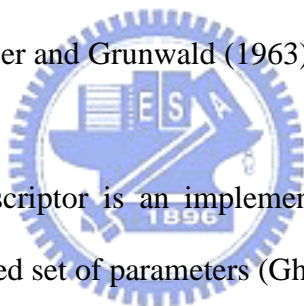
**Energy:** The energy of the selected compound.

**Chiral centers:** Number of chiral center (R or S) in a molecule.

**AlogP and molar refractivity (MolRef):** LogP (the octanol/water partition coefficient) and molar refractivity are molecular descriptors that can be used to relate chemical structure to observed chemical behavior. LogP is related to the hydrophobic character of the molecule. The molecular refractivity index of a substituent is a combined measure of its size and polarizability.

The QSAR descriptor ALogP and molar refractivity are calculated using the method described by Ghose & Crippen (1989). In this atom-based approach, each atom of the molecule is assigned to a particular class, with additive contributions to the total value of logP and molar refractivity.

For more information, see Leffler and Grunwald (1963).



**AlogP98:** The AlogP98 descriptor is an implementation of the atom-type-based AlogP method using the latest published set of parameters (Ghose et al. 1998).

**Desolvation free energy for water ( $F_{H_2O}$ ) and octanol ( $F_{oct}$ ):**  $F_{oct}$  and  $F_{H_2O}$  are physiochemical properties associated with LFE models of a molecule. These properties have proven useful as molecular descriptors in structure–activity analyses. All LFE computations are based solely on the connectivity of the atoms in a molecule. LFE computations are not conformationally dependent.

$F_{oct}$  is the 1-octanol desolvation free energy and  $F_{H_2O}$  is the aqueous desolvation free energy derived from a hydration shell model developed by Hopfinger, where  $F_{oct}$  and  $F_{H_2O}$  are in kcal mol<sup>-1</sup>.

For more information, see Hopfinger (1973; 1980) Pearlman (1980).

**Heat of formation (Hf):** The enthalpy for forming a molecule from its constituent atoms, a measure of the relative thermal stability of a molecule. This descriptor is calculated using the MNDO semi-empirical molecular orbital method of Dewar. MNDO is the most rigorous quantum-chemical technique available in QSAR+ and has a wide range of applicability in conformational analysis, intermolecular modeling, and chemical reaction modeling. The atom limit of MNDO is 300 atoms or 300 atomic orbitals (whichever is less) per molecule. The atoms treated by MNDO are: H, B, C, N, O, F, Al, Si, P, S, and Cl. For more information, see Dewar and Thiele (1977a; 1977b).

## 2.4 Build LigSeeSVM prediction Model

### A. Divide data set into training set and testing set

To choose the known ligands used as the positive examples of the SVM training set, we calculated the molecular similarity. Euclidean distance was used as the measure of similarity in the multidimensional descriptor space. The latter distance  $d_{ij}$  between any two compounds  $i$  and  $j$  in  $N$ -dimensional descriptor space was calculated as

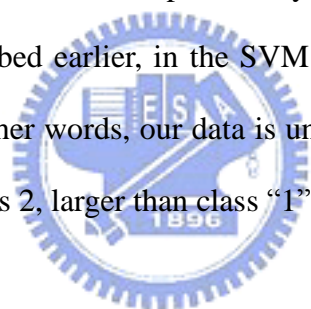
$$d_{ij} = \sqrt{\sum_{n=1}^N (X_{in} - X_{jn})^2}$$

where  $X_{in}$  and  $X_{jn}$  are the values of  $n$ th descriptor for compounds  $i$  and  $j$ , respectively, and the summation is over all descriptors [38]. Based on UPGMA algorithm [39], we created a rooted tree of the known ligands. Compounds with the maximum distance (least similarity) were chosen as the positive examples of the SVM training set combined with randomly chosen molecules from 990 ACD molecules or 7300 CMC molecules as the negative examples. In the SVM training set, the ratio of the number of negative and positive data is about 20. For example, in the ACD screening set, the SVM training set consisted of two or three known

ligands as positive examples and 40 randomly chosen compounds from 990 ACD screening set as negative examples. The remained known ligands and 950 ACD molecules made up the SVM testing set. In the CMC screening set, we chose half of the known ligands as positive examples and 200 randomly chosen compounds from 7300 CMC screening set as negative examples to make up the SVM training set, and the SVM testing set was composed of the remained half known ligands and 7100 CMC molecules.

### **B. Add $-b$ and $-wi$ parameters**

When training a SVM model, add  $-b$  parameter allows LibSVM to train a model for probability estimates. Next, we transfer the probability score to Z-score and sort according to Z-score. Besides, as we described earlier, in the SVM training set, the ratio of negative and positive data is about 20. In other words, our data is unbalanced, so we add  $-wi$  parameter to give the penalty for class “-1” is 2, larger than class “1”.



### **C. Rank combination**

In this study, we combined the results of SVM-AP and SVM-PC using a “rank data fusion” method in which a new rank is generated by sorting the average rank of SVM-AP and SVM-PC. Application of this hybrid method to each test set to improve the performance of both SVM-AP and SVM-PC predicted models.

## Chapter 3

### Evaluation LigSeeSVM on ACD Database

We have made seven screening sets against different target proteins, thymidine kinase (TK), estrogen receptor with antagonists (ER-antagonists), estrogen receptor with agonists (ER-agonists), serotonin receptor, muscarinic receptor, histamine receptor, and GABA<sub>A</sub> receptor. Each screening set includes several known active ligands and 990 ACD randomly selected compounds. Five common metrics were used to evaluate the screening quality, including the true hit (the percentage of active ligands retrieved from the database), hit rate (the percentage of active ligands in the hit list), goodness-of-hit (GH), ROC curves, and false positive rate.



### 3.1 Thymidine Kinase

#### A. Preparation of screening data set

To evaluate the virtual screening utility of SVM, we used the thymidine kinase benchmark from Bissantz et al. [26]. It consisted of 10 known ligands (Figure 4) and 990 compounds ACD screening set mentioned above. We chose two known ligands (1kim.THM and 2ki5.AC2) based on the UPGMA rooted tree (Figure 5A) as positive examples of SVM training set and 40 randomly chosen compounds from 990 ACD screening set as negative examples. The SVM testing set including 8 remained known ligands and 950 ACD molecules.

#### B. Virtual screening of TK substrates



Table 5 shows the result of the remaining 8 TK known ligands in SVM testing set using different descriptors. The results of different descriptors are also shown in Figure 5.

Some common factors were used to evaluate the screening quality, including coverage (the percentage of active ligands retrieved from the database), yield (the percentage of active ligands in the hit list), false-positive (FP) rate, enrichment, and goodness-of-hit (GH). The coverage (true positive rate) is defined as  $A_h/A$  (%),  $A_h/Th$  (%) is the yield (hit rate), and the FP rate is defined as  $(Th - Ah)/(T - A)$  (%). The enrichment is defined as  $(Ah/Th)/(A/T)$ .  $A_h$  is the number of active ligands among the  $Th$  highest ranking compounds, which is called the hit list,  $A$  is the total number of active ligands in the database, and  $T$  is the total number of compounds in the database. The GH score is defined as [40]

$$GH = \left( \frac{A_h(3A + T_h)}{4T_h A} \right) / \left( 1 - \frac{T_h - A_h}{T - A} \right)$$

In the case of TK  $A$  and  $T$  are 10 and 1000, respectively.

The main objective of this study was to evaluate the behavior of different predicted models in virtual screening. Figure 6 shows the results of SVM-AP, SVM-PC, and LigSeeSVM. We tested SVM with different descriptors to evaluate the performance and the search behavior. The screening quality generally improved in the case of LigSeeSVM. As shown in Figure 6A, the hit rates for different predicted models are 50% (SVM-AP), 40% (SVM-PC), and 72.7% (LigSeeSVM) when the TP rate is 100%. In the case of LigSeeSVM and the TP rate is 100%, the GH score is 0.79 (Figure 6B) and the FP rate is 0.3% (Table 6).

Table 6 shows the comparative false positive rates on the same target protein and screening database at true positive rates 80% and 100% (Surflex [3]). For the true positive rate of 100%, the FP rate for LigSeeSVM is 0.3%. Besides, the FP rates for Surflex is 0.7%, SVM-AP is 0.8%, and SVM-PC is 1.3%. The performance of LigSeeSVM is better than Surflex.

## 3.2 Estrogen Receptor

### A. Preparation of screening data set

We have applied SVM to virtual screening against ER  $\alpha$  with a testing sets composed of 11 known antagonists of ER  $\alpha$  (Figure 7), 10 known agonists of ER  $\alpha$  (Figure 8) [29] and 990 randomly selected compounds from ACD (Available Chemicals Directory). As we described earlier, based on the UPGMA rooted tree (Figure 5B, Figure 5C), we chose two known ER  $\alpha$  antagonists (EST02 and EST11) as positive examples of the SVM training set and 40 randomly chosen molecules from 990 ACD screening set, and the SVM testing set was composed of the remained 9 known antagonists and 950 ACD compounds. In the ER  $\alpha$  agonists case, we chose three known ER  $\alpha$  agonists (ESA01, ESA04, and ESA09) as positive examples of the SVM training set and 40 randomly chosen molecules from 990 ACD screening set, and the SVM testing set including remained 7 known agonists and 950 ACD molecules.

### B. Virtual screening of ER $\alpha$ antagonists and agonists

Table 7 shows the result of the ER  $\alpha$  antagonists and Table 9 shows the result of ER  $\alpha$  agonists in SVM testing set using different predicted models. The results of different predicted models are also shown in Figure 9(ER  $\alpha$  antagonists). and Figure 10(ER  $\alpha$  agonists). As shown in Figure 9A, the hit rates of ER  $\alpha$  antagonists for different predicted models are 23% (SVM-AP), 17.3% (SVM-PC), and 69.2% (LigSeeSVM) when the TP rate is 100%. In the case of LigSeeSVM and the TP rate is 100%, the GH score is 0.77 (Figure 9B) and the FP rate is 0.6% (Table 8). Table 8 shows the comparative false positive rates on the same target protein and screening database at true positive rates 80% and 100%. For the true

positive rate of 100%, the FP rate for LigSeeSVM is 0.6%. Besides, the FP rates for Surflex is 13.4%, SVM-AP is 3.2%, and SVM-PC is 4.5%. The performance of LigSeeSVM is much better than Surflex. For the ER  $\alpha$  agonists case, as shown in Figure 10, the hit rates for different predicted models are 15.9% (SVM-AP), 28% (SVM-PC), and 100% (LigSeeSVM) when the TP rate is 100%. Table 10, 11 shows top ranked 10 compound structures in the case of ER  $\alpha$  agonists.

### 3.3 GPCR and GABA<sub>A</sub> Receptor

#### A. Preparation of screening data set

To compare with Surflex-Sim [3], we used the same molecules (Figure 2) which described in the paper as positive examples of SVM training set and combined with 40 randomly chosen compounds from 990 ACD screening set as negative examples of training set. A and B are serotonin ligands. C (tolterodine), D, and E are muscarinic antagonists. Molecules F–H (bromodiphenhydramine, pyrilamine, and azatadine) are H1 receptor antagonists. Molecules I–K are GABA<sub>A</sub> receptor agonists (diazepam, alprazolam, and zopiclone). The same molecules (Table 1) which described in the paper [3] also used to test the SVM models. The serotonin SVM testing set including 29 known serotonin ligands and 950 ACD molecules. The muscarinic SVM testing set was composed of 43 known muscarinic ligands (Table 1) and 950 ACD molecules. Again, the histamine SVM testing set consisted of 49 known histamine ligands (Table 1) and 950 ACD molecules. The GABA<sub>A</sub> SVM testing set including 15 known GABA<sub>A</sub> ligands (Table 1) and 950 ACD molecules.

#### B. Virtual screening of GPCR and GABA<sub>A</sub>

Figure 11A shows the ROC curve of muscarinic using SVM, and Figure 11B shows the ROC curve of muscarinic using Surfex-Sim. For the muscarinic case, the FP rate of LigSeeSVM is 0.21 when the TP rate is 1. Figure 12A shows the ROC curve of histamine using SVM, and Figure 12B shows the ROC curve of histamine using Surfex-Sim. For the histamine case, the FP rate of LigSeeSVM is 0.14 when the TP rate is 1. Figure 13A shows the ROC curve of GABA<sub>A</sub> using SVM, and Figure 13B shows the ROC curve of GABA<sub>A</sub> using Surfex-Sim. For the GABA<sub>A</sub> case, the FP rate of LigSeeSVM is 0.02 when the TP rate is 1. Figure 14A shows the ROC curve of serotonin using SVM, and Figure 14B shows the ROC curve of serotonin using Surfex-Sim. For the serotonin case, which was built on just two ligands of vastly different structure than the test ligands, the FP rate of LigSeeSVM is 0.29 higher than other three cases when the TP rate is 1.



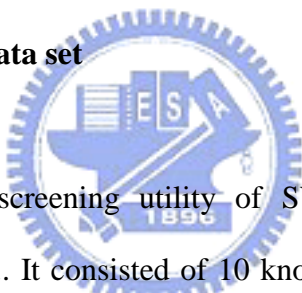
## Chapter 4

### Evaluation LibSVM on CMC Database

We have made seven screening sets against different target proteins, thymidine kinase (TK), estrogen receptor with antagonists (ER-antagonists), estrogen receptor with agonists (ER-agonists), serotonin receptor, muscarinic receptor, histamine receptor, and GABA<sub>A</sub> receptor. Each screening set includes the same known active ligands as we described earlier and 7300 compounds from CMC.

#### 4.1 Thymidine Kinase

##### A. Preparation of screening data set



To evaluate the virtual screening utility of SVM, we used the thymidine kinase benchmark from Bissantz et al.. It consisted of 10 known ligands and 990 compounds ACD screening set mentioned above. We used the same 10 known ligands but the 7300 compounds from CMC displaced the 990 ACD compounds. Based on the UPGMA rooted tree (Figure 5A), we chose 5 known ligands (1kim.THM, 1e2m.HPT, 2ki5.AC2, 1ki2.GA2, 1ki3.PE2) as positive examples of SVM training set and 200 randomly chosen compounds from 7300 CMC compounds as negative examples. The SVM testing set including 5 remained known ligands and 7100 CMC molecules.

##### B. Virtual screening of TK substrates

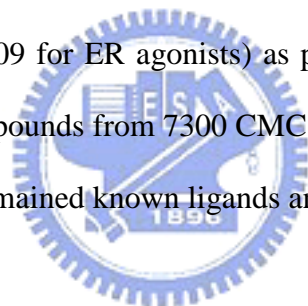
Table 12 shows that the remaining 5 known ligands in which they ranked in a screen including 7100 randomly chosen compounds from CMC. The ranks of the 5 known ligands

were: 1, 2, 3, 4, 5. Table 13 shows the top ranked 10 compounds' structures in the case of LigSeeSVM.

## 4.2 Estrogen Receptor

### A. Preparation of screening data set

We have applied SVM to virtual screening against ER  $\alpha$  with a testing sets composed of 11 known antagonists of ER  $\alpha$ , 10 known agonists of ER  $\alpha$  and 7300 randomly selected compounds from CMC. Based on the UPGMA rooted tree (Figure 5B, Figure 5C), we chose 5 known ligands (EST01, EST02, EST03, EST08, EST09 for ER antagonists, and ESA01, ESA03, ESA04, ESA08, ESA09 for ER agonists) as positive examples of SVM training set and 200 randomly chosen compounds from 7300 CMC compounds as negative examples. The SVM testing set including 5 remained known ligands and 7100 CMC molecules.



### B. Virtual screening of ER $\alpha$ antagonists and agonists

Table 14 shows that the remaining 6 known ER  $\alpha$  antagonists in which they ranked in a screen including 7100 randomly chosen compounds from CMC. The ranks of the 6 known antagonists were: 1, 2, 3, 4, 5, 7. Table 15 shows the top ranked 10 compounds' structures in the case of LigSeeSVM. Table 16 shows that the remaining 5 known ER  $\alpha$  agonists in which they ranked in a screen including 7100 randomly chosen compounds from CMC. The ranks of the 5 known agonists were: 1, 2, 3, 4, 5. Table 17 shows the top ranked 10 compounds' structures in the case of LigSeeSVM.

### 4.3 GPCR and GABA<sub>A</sub> Receptor

#### A. Preparation of screening data set

Based on the screening utility of SVM described above, we applied SVM to virtual screening for GPCR and GABA<sub>A</sub> with a screening set including 100 known ligands (Table 1) and 7300 molecules from the CMC. According to the UPGMA rooted tree, we chose half known ligands as positive examples of SVM training set and 200 randomly chosen compounds from 7300 CMC compounds as negative examples. The SVM testing set including remained known ligands and 7100 CMC molecules.

#### B. Virtual screening of GPCR and GABA<sub>A</sub> ligands

For the serotonin case, Table 18 shows that the top ranked 10 compounds structures of the LigSeeSVM in a screen including 7100 randomly chosen compounds from the CMC. For the muscarinic case, Table 19 shows that the top ranked 10 compounds structures of the LigSeeSVM in a screen including 7100 randomly chosen compounds from the CMC. Again, for the histamine case, Table 20 shows that the top 10 ranked compounds structures of the LigSeeSVM in a screen including 7100 randomly chosen compounds from the CMC. For the GABA<sub>A</sub> case, Table 21 shows that the top ranked 10 compounds structures of the LigSeeSVM in a screen including 7100 randomly chosen compounds from the CMC.

## Chapter 5

### Conclusions

#### 5.1 Summary

In summary, the results presented here suggests that the SVM approach is well suited to the drug screening problem and yields a prediction accuracy superior to the earlier study derived from Surfex-Sim. The majority compounds with high Z-score have structures similar to the known ligands. Some compounds with high Z-score have different structures compared with the known ligands, and these compounds have more possibility to become the novel, potential lead compounds. Moreover, our study also shows that the strategy of rank combination have great improved the results of drug screening.

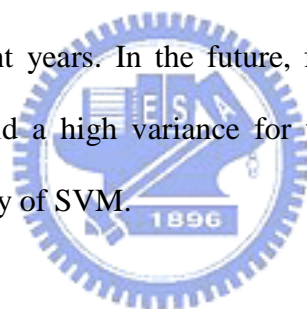
#### 5.2 Major Contributions and Future Perspectives

To apply LibSVM to virtual screening, a suitable compound descriptor is essential. The major molecule descriptors used in this work are derived from 2D molecular topology (825 different atom pair descriptors). The major disadvantages of 2D ligand-based drug screening are 2D methods cannot describe compound physicochemical properties and distinguish stereoisomers. By integrating a number of Accelrys Cerius<sup>2</sup> QSAR module default and thermodynamic descriptors, we improved the disadvantages of 2D ligand-based drug screening. The virtual screening utility of SVM has been evaluated by 14 screening sets including TK substrates, ER antagonists, ER agonists, GABA<sub>A</sub> ligands, three kinds of GPCR ligands combined with ACD or CMC databases. The screening performance of SVM is superior to the other public ligand-based approach.



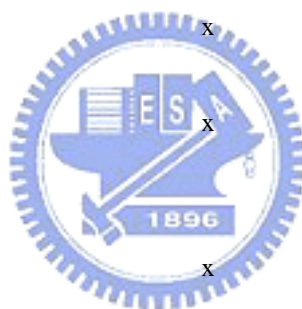
Earlier in this lab, we have applied GEMDOCK to virtual screening for the envelope protein of dengue virus to screen potential inhibitors from the CMC. The nine candidates we recommended have been tested by biological experiments (cooperation with Dr. Yun-Lian Yang) and we found that MCMC00007079 could suppress the activity of dengue virus with concentrations of 1 mM and 10 mM. Using this compound as a training model of SVM to screen other potential inhibitors from the chemical database will be a good start to refine them by lead optimization.

The use of descriptors unrelated to a particular type of properties or biological activity likely generates noise in a statistical learning system, which may affect the prediction accuracy of that system. Thus, the issue of feature selection in the SVM framework has received attention in the recent years. In the future, feature selection approach will be the point to be developed to avoid a high variance for unseen molecules (overfitting) and to improve the prediction accuracy of SVM.



**Table 1.** 100 positive compounds of GPCR and GABA<sub>A</sub> studied in this thesis, with annotation of known targets [3].

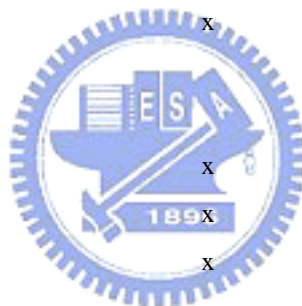
Molecule name	GPCR			GABA <sub>A</sub>
	Serotonin receptor ligand	Muscarinic receptor ligand	Histamine receptor ligand	GABA <sub>A</sub> receptor ligand
abecarnil				x
alose tron	x			
alpidem				x
amitriptyline	x	x	x	
amoxapine		x	x	
astemizole			x	
atropine		x		
bcce <sup>a</sup>				x
benztropine		x	x	
bethanechol		x		
bromolysergide	x			
brompheniramine		x	x	
bupropion			x	
carbachol		x		
carbinoxamine			x	
cetirizine			x	
chlorpheniramine		x	x	
chlorpromazine	x	x	x	
clemastine			x	
clobazam				x
clomipramine		x	x	
clozapine	x	x	x	
cocaine		x		
cyproheptadine		x	x	
darifenacin		x		
desipramine		x	x	
dicyclomine		x		
dolasetron	x			
dotarizine	x			
doxepin		x	x	
estazolam				x
fluphenazine		x	x	
flutoprazepam				x



granisteron	x			
halazepam				x
haloperidol		x		
hydroxyzine			x	
iloperidone	x			
imipramine		x	x	
itasetron	x			
ketanserin	x			
levocabastine			x	
lidocaine		x		
loratadine			x	
loxapine		x	x	
maprotiline		x	x	
meclizine			x	
mesoridazine		x	x	
metergoline	x			
methotrimeprazine			x	
methscopolamine			x	
methysergide	x			
metitepine	x		x	
mianserin	x		x	
midazolam				x
molindone		x		
nefazodone	x			
nortriptyline	x	x	x	
olanzapine	x			
ondansetron	x			
oxetorone	x			
oxybutynin		x		
perospirone	x			
perphenazine		x	x	
phenindamine			x	
phenoxybenzamine			x	
pilocarpine		x		
pimozide		x		
pindolol	x			
pizotyline	x			
prazepam				x



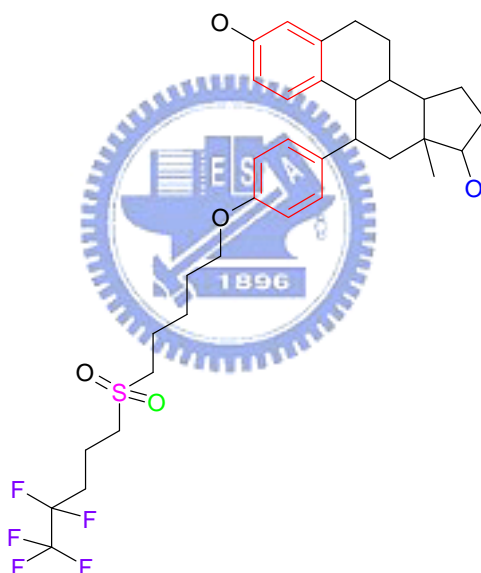
prochlorperazine		x		x	
procyclidine		x			
promethazine		x		x	
protriptyline		x		x	
quazepam					x
ramosteron	x				
risperidone	x			x	
ritanserin	x			x	
sertindole	x	x		x	
suriclone					x
telenzipine		x			
terfenadine				x	
tetrazepam					x
thiethylperazine				x	
thioridazine		x		x	
thiothixene		x		x	
tiotropium		x			
trazodone				x	
triazolam					x
trifluoperazine		x		x	
triflupromazine		x		x	
trimipramine		x		x	
tripeleennamine				x	
triprolidine				x	
tropisetron	x				
zaleplon					x
ziprasidone	x				
zolpidem					x
zotepine	x				



<sup>a</sup>Note: bcce Is Ethyl  $\beta$ -Carboline-3-carboxylate

**Table 2.** Ten atom types used for atom pair descriptors [11].

	Atom type	Description	Atom type mol2 format
1	C.ar	aromatic carbons	C.ar
2	C.na	nonaromatic carbons	C.3, C.2, C.1, and C.cat
3	N.ar	aromatic nitrogen	N.ar
4	N.na	nonaromatic nitrogen	N.3, N.2, N.1, N.am, N.4, and N.p1
5	O.3	oxygen atoms in the sp <sup>3</sup> hybridization state	O.3
6	O.2	oxygen atoms in the sp <sup>2</sup> hybridization state	O.2
7	S	all sulfur atoms	S.3, S.2, S.O, and S.O2
8	P.3	phosphorus atoms	P.3
9	X	halogen atoms	F, Cl, Br, and I
10	other atoms		The other atom types



**Table 3.** Six thermodynamic descriptors available in Accelrys Cerius<sup>2</sup> QSAR+ module [13]

Symbol	Description
AlogP	Log of the partition coefficient
AlogP98	Log of the partition coefficient, atom-type value
Fh2o	Desolvation free energy for water
Foct	Desolvation free energy for octanol
Hf	Heat of formation
MolRef	Molar refractivity



**Table 4.** 13 default descriptors available in Accelrys Cerius<sup>2</sup> QSAR+ module [13]

Symbol	Description
Apol	Sum of atomic polarizabilities
Dipole-mag	The strength and orientation behavior of a molecular in an electrostatic field
RadOfGyration	The radius of gyration
Area	Molecular surface area
MW	Molecular weight
Vm	Molecular volumn
Density	Density
PMI-mag	Principal moments of inertia
Rotlbonds	Number of rotatable
Hbond acceptor	Number of hydrogen-bond acceptors
Hbond donor	Number of hydrogen-bond donors
Energy	The energy of the compound
Chiral centers	Number of chiral center (R or S) in a molecule



**Table 5.** Accuracies of three kinds SVM models for 8 TK test substrates.

name	thymidine kinase substrate					
	SVM-AP		SVM-PC		LigSeeSVM	
	Z-score	rank	Z-score	rank	score <sup>a</sup>	rank <sup>b</sup>
1e2k	8.10	3	7.04	9	6	2
1e2m	6.51	11	5.39	17	14	7
1e2n	3.89	16	4.71	20	18	11
1e2p	6.23	13	7.88	4	8.5	5
1ki2	8.72	1	7.92	2	1.5	1
1ki3	7.62	9	7.88	5	7	4
1ki6	6.16	14	6.44	11	12.5	6
1ki7	8.04	4	7.43	8	6	3

<sup>a</sup> the average rank of SVM-AP and SVM-PC

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC





**Table 6.** Comparison our three SVM models with Surfex-Sim on false positive rates for thymidine kinase substrates and ACD database at true positive rates 80% and 100%.

TP, %	false positives from random ligands, %			
	thymidine kinase substrate			
	Surflex-Sim <sup>a</sup>	SVM-AP	SVM-PC	LigSeeSVM <sup>b</sup>
80	0.3	0.7	0.5	0.0
100	0.7	0.8	1.3	0.3

<sup>a</sup> results directly summarized from [3]

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC



**Table 7.** Accuracies of three kinds SVM models for 9 ER test antagonists.

name	estrogen receptor antagonist					
	SVM-AP		SVM-PC		LigSeeSVM	
	Z-score	rank	Z-score	rank	score <sup>a</sup>	rank <sup>b</sup>
EST01	5.41	14	4.14	17	15.5	7
EST03	2.06	39	5.74	3	21	11
EST04	5.72	7	2.58	39	23	12
EST05	5.67	9	5.77	2	5.5	3
EST06	5.72	4	5.25	6	5	1
EST07	5.72	5	5.49	5	5	2
EST08	5.56	13	1.90	52	32.5	15
EST09	4.36	21	5.83	1	11	5
EST10	5.67	10	4.11	18	14	6

<sup>a</sup> the average rank of SVM-AP and SVM-PC

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC



**Table 8.** Comparison our three SVM models with Surfex-Sim on false positive rates for estrogen receptor antagonists and ACD database at true positive rates 80% and 100%.

TP, %	false positives from random ligands, %			
	estrogen receptor antagonist			
	Surflex-Sim <sup>a</sup>	SVM-AP	SVM-PC	LigSeeSVM <sup>b</sup>
80	0.0	0.7	1.2	0.4
100	13.4	3.2	4.5	0.6

<sup>a</sup> results directly summarized from [3]

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC



**Table 9.** Accuracies of three kinds SVM models for 7 ER test agonists.

name	estrogen receptor agonist					
	SVM-AP		SVM-PC		LigSeeSVM	
	Z-score	rank	Z-score	rank	score <sup>a</sup>	rank <sup>b</sup>
ESA02	8.12	2	13.71	3	2.5	1
ESA03	1.40	44	8.66	5	24.5	7
ESA05	9.18	1	0.28	25	13	5
ESA06	8.67	4	13.88	1	2.5	2
ESA07	8.04	6	13.86	2	4	3
ESA08	1.46	38	8.37	6	22	6
ESA10	6.30	7	13.70	4	5.5	4

<sup>a</sup> the average rank of SVM-AP and SVM-PC

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC



**Table 10.** Top ranked 10 compounds of our three SVM models on screening ER agonists from the ACD chemical database.

SVM-AP		SVM-PC		LigSeeSVM	
ID	Structure	ID	Structure	ID	Structure
ESA05		ESA06		ESA02	
ESA02		ESA07		ESA06	
MFCD00003689		ESA02		ESA07	
ESA06		ESA10		ESA10	
MFCD00010489		ESA03		ESA05	
ESA07		ESA08		ESA08	
ESA10		MFCD00000439		ESA03	
MFCD00002206		MFCD00011526		MFCD00010480	
MFCD00012180		MFCD00009902		MFCD00003650	
MFCD00005048		MFCD00003747		MFCD00003678	

**Table 11.** The remaining 5 TK known ligands in which they ranked in a screen including 7100 randomly chosen compounds from CMC.

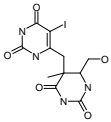
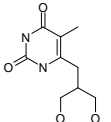
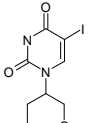
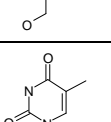
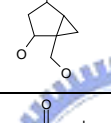
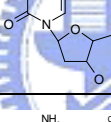
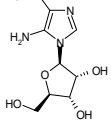
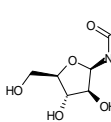
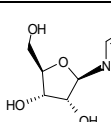
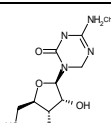
name	thymidine kinase substrate					
	SVM-AP		SVM-PC		LigSeeSVM	
	Z-score	rank	Z-score	rank	score <sup>a</sup>	rank <sup>b</sup>
1e2n	15.24	1	27.91	2	1.5	1
1e2p	14.59	3	30.89	1	2	2
1ki6	14.79	2	18.46	4	3	3
1e2k	11.80	6	27.61	3	4.5	4
1ki7	14.55	4	17.39	6	5	5

<sup>a</sup> the average rank of SVM-AP and SVM-PC

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC



**Table 12.** Top ranked 10 compounds of the LigSeeSVM on screening TK ligands from the CMC chemical database.

Rank	ID	Structure	Z-score <sup>a</sup>
1	1e2n_hmtt		21.57
2	1e2p_dhbt		22.74
3	1ki6_ahiu		16.63
4	1e2k_mct		19.70
5	1ki7_idu		15.97
6	MCMC00005826		8.58
7	MCMC00004760		7.00
8	MCMC00000973		6.82
9	MCMC00004622		6.57
10	MCMC00007639		5.39

<sup>a</sup> the average Z-score of SVM-AP and SVM-PC

**Table 13.** The remaining 6 ER known antagonists in which they ranked in a screen including 7100 randomly chosen compounds from CMC.

name	estrogen receptor antagonist					
	SVM-AP		SVM-PC		LigSeeSVM	
	Z-score	rank	Z-score	rank	score <sup>a</sup>	rank <sup>b</sup>
EST04	10.47	11	17.92	4	7.5	5
EST05	10.83	7	19.12	2	4.5	3
EST06	11.33	1	18.88	3	2	2
EST07	11.03	5	17.88	5	5	4
EST10	11.19	2	19.30	1	1.5	1
EST11	11.00	6	11.01	15	10.5	7

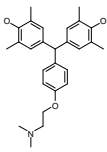
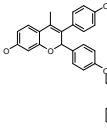
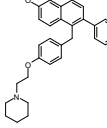
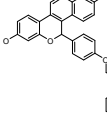
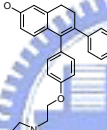
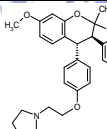
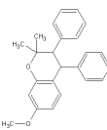
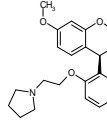
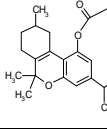
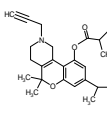
<sup>a</sup> the average rank of SVM-AP and SVM-PC

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC





**Table 14.** Top ranked 10 compounds of the LigSeeSVM on screening ER antagonists from the CMC chemical database.

Rank	ID	Structure	Z-score <sup>a</sup>
1	EST10		15.25
2	EST06		15.10
3	EST05		14.97
4	EST07		14.45
5	EST04		14.20
6	MCMC00006676		11.21
7	EST11		11.01
8	MCMC00007662		12.62
9	MCMC00004449		11.66
10	MCMC00005381		12.94

<sup>a</sup> the average Z-score of SVM-AP and SVM-PC

**Table 15.** The remaining 5 ER known agonists in which they ranked in a screen including 7100 randomly chosen compounds from CMC.

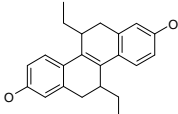
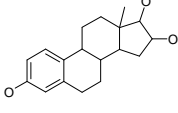
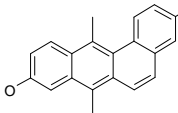
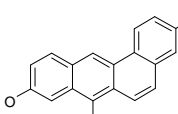
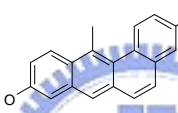
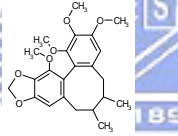
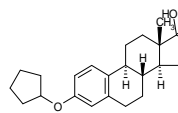
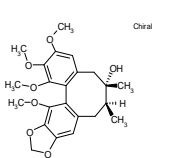
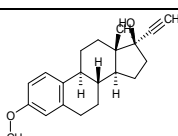
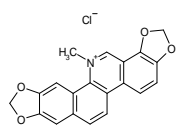
name	estrogen receptor agonist					
	SVM-AP		SVM-PC		LigSeeSVM	
	Z-score	rank	Z-score	rank	score <sup>a</sup>	rank <sup>b</sup>
ESA02	23.14	3	17.93	1	2	1
ESA05	23.37	2	16.81	5	3.5	2
ESA06	22.69	4	17.43	3	3.5	3
ESA10	22.66	5	17.68	2	3.5	4
ESA07	22.21	6	17.23	4	5	5

<sup>a</sup> the average rank of SVM-AP and SVM-PC

<sup>b</sup> sort according to the average rank of SVM-AP and SVM-PC

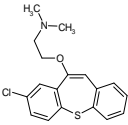
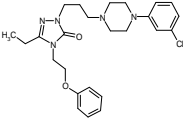


**Table 16.** Top ranked 10 compounds of the LigSeeSVM on screening ER agonists from the CMC chemical database.

Rank	ID	Structure	Z-score <sup>a</sup>
1	ESA02		20.53
2	ESA05		20.09
3	ESA06		20.06
4	ESA10		20.17
5	ESA07		19.72
6	MCMC00010094		7.18
7	MCMC00007029		5.35
8	MCMC00006241		4.70
9	MCMC00006985		4.23
10	MCMC00006197		6.12

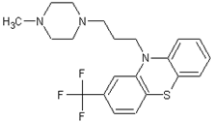
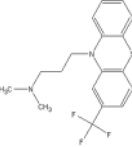
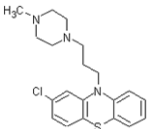
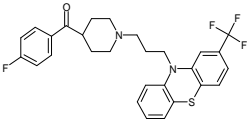
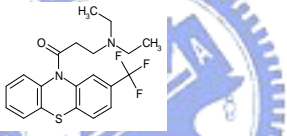
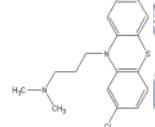
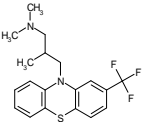
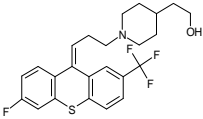
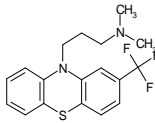
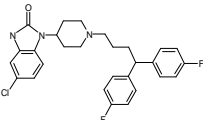
<sup>a</sup> the average Z-score of SVM-AP and SVM-PC

**Table 17.** Top ranked 10 compounds of the LigSeeSVM on screening serotonin receptor ligands from the CMC chemical database.

Rank	ID	Structure	Z-score <sup>a</sup>
1	chlorpromazine		6.05
2	MCMC00005489		5.97
3	tropisetron		5.80
4	MCMC00003399		5.75
5	MCMC00000657		5.80
6	MCMC00002871		5.66
7	MCMC00002701		5.60
8	MCMC00009837		5.53
9	MCMC00000030		5.66
10	MCMC00005368		5.53

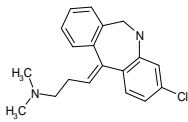
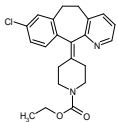
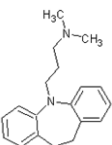
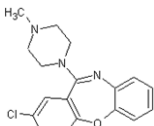
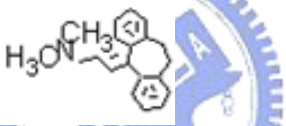
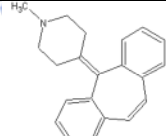
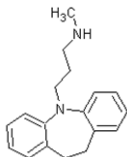
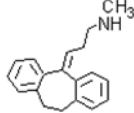
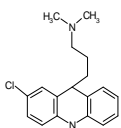
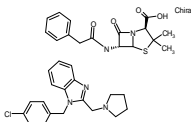
<sup>a</sup> the average Z-score of SVM-AP and SVM-PC

**Table 18.** Top ranked 10 compounds of the LigSeeSVM on screening muscarinic receptor ligands from the CMC chemical database.

Rank	ID	Structure	Z-score <sup>a</sup>
1	trifluoperazine		7.73
2	triflupromazine		7.00
3	prochlorperazine		7.33
4	MCMC00004614		6.49
5	MCMC00003532		5.64
6	chlorpromazine		6.59
7	MCMC00001866		5.25
8	MCMC00004241		5.12
9	MCMC00000861		4.96
10	MCMC00004142		5.97

<sup>a</sup> the average Z-score of SVM-AP and SVM-PC

**Table 19.** Top ranked 10 compounds of the LigSeeSVM on screening histamine receptor ligands from the CMC chemical database.

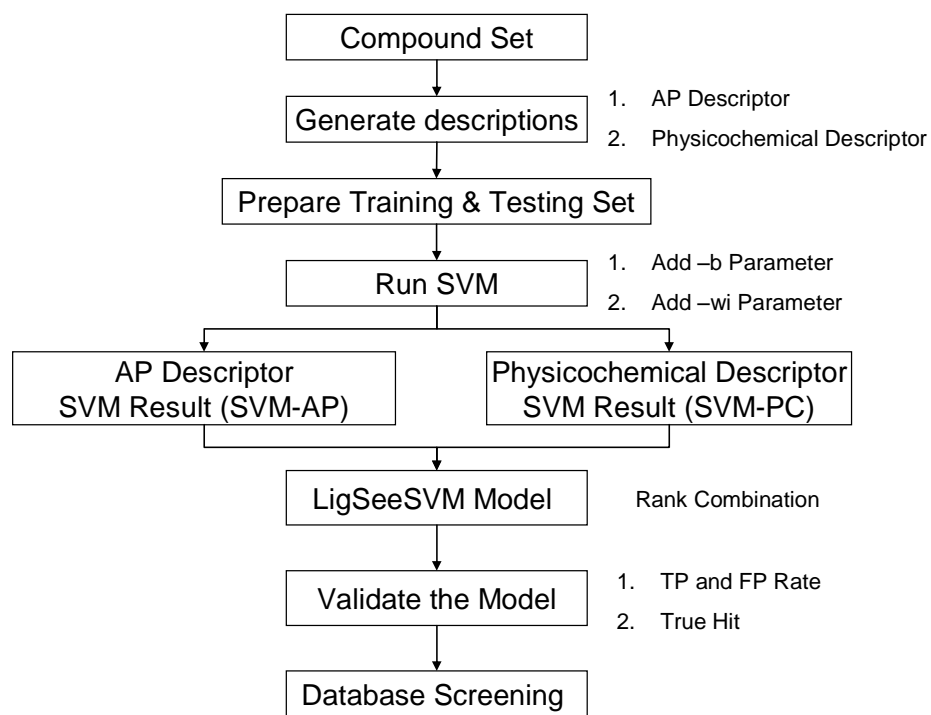
Rank	ID	Structure	Z-score <sup>a</sup>
1	MCMC00002400		5.38
2	MCMC00005251		5.56
3	imipramine		5.22
4	loxapine		5.19
5	amitriptyline		5.32
6	cypheptadine		5.17
7	desipramine		5.02
8	nortriptyline		5.29
9	MCMC00002252		4.77
10	MCMC00007349		5.42

<sup>a</sup> the average Z-score of SVM-AP and SVM-PC

**Table 20.** Top ranked 10 compounds of the LigSeeSVM on screening GABA<sub>A</sub> receptor ligands from the CMC chemical database.

Rank	ID	Structure	Z-score <sup>a</sup>
1	MCMC00005351		22.97
2	MCMC00006054		19.88
3	triazolam		10.98
4	MCMC00003492		16.41
5	MCMC00002542		9.71
6	MCMC00003781		17.32
7	MCMC00003527		16.59
8	midazolam		6.77
9	MCMC00004114		6.39
10	MCMC00003490		7.86

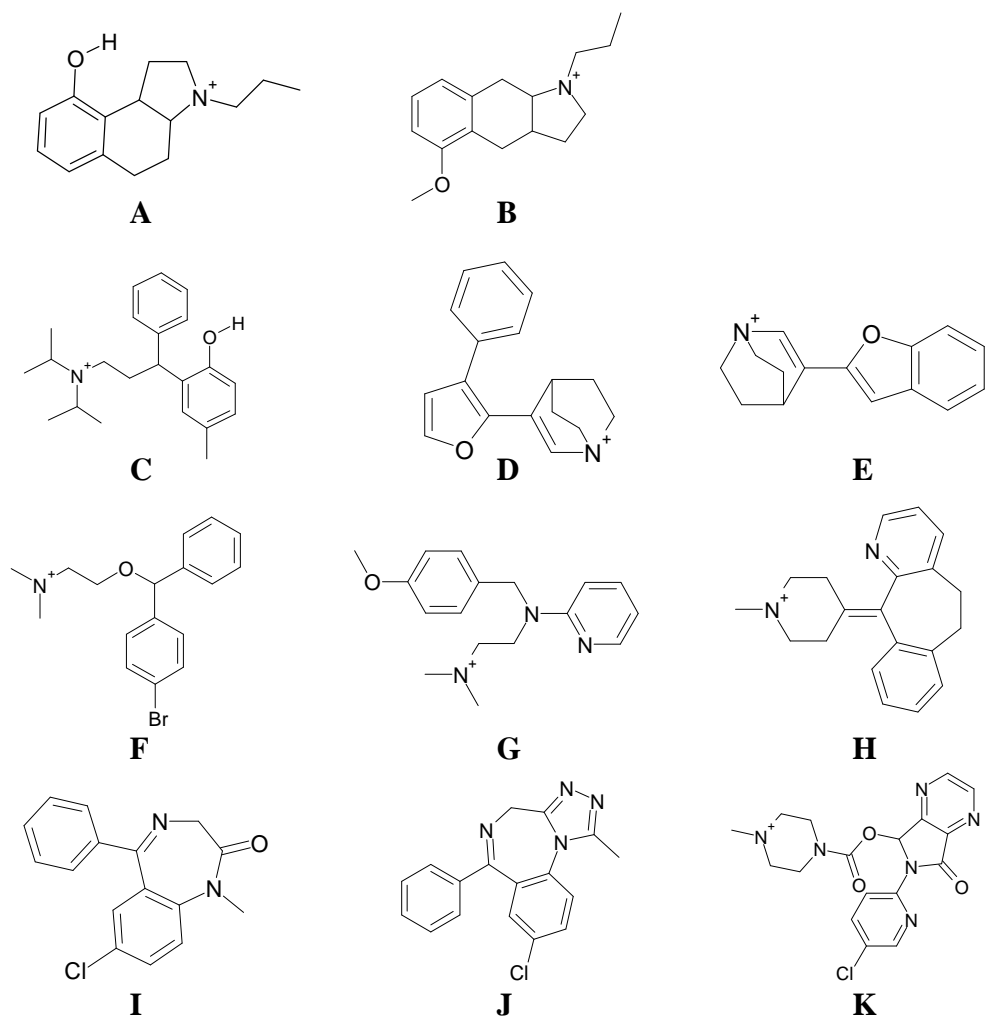
<sup>a</sup> the average Z-score of SVM-AP and SVM-PC



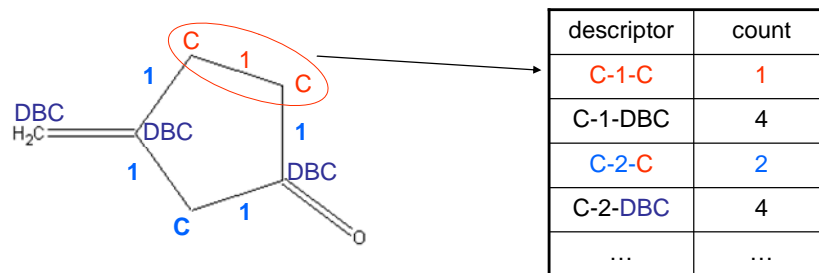
**Figure 1.** Overview of our method for ligand-based compound screening.





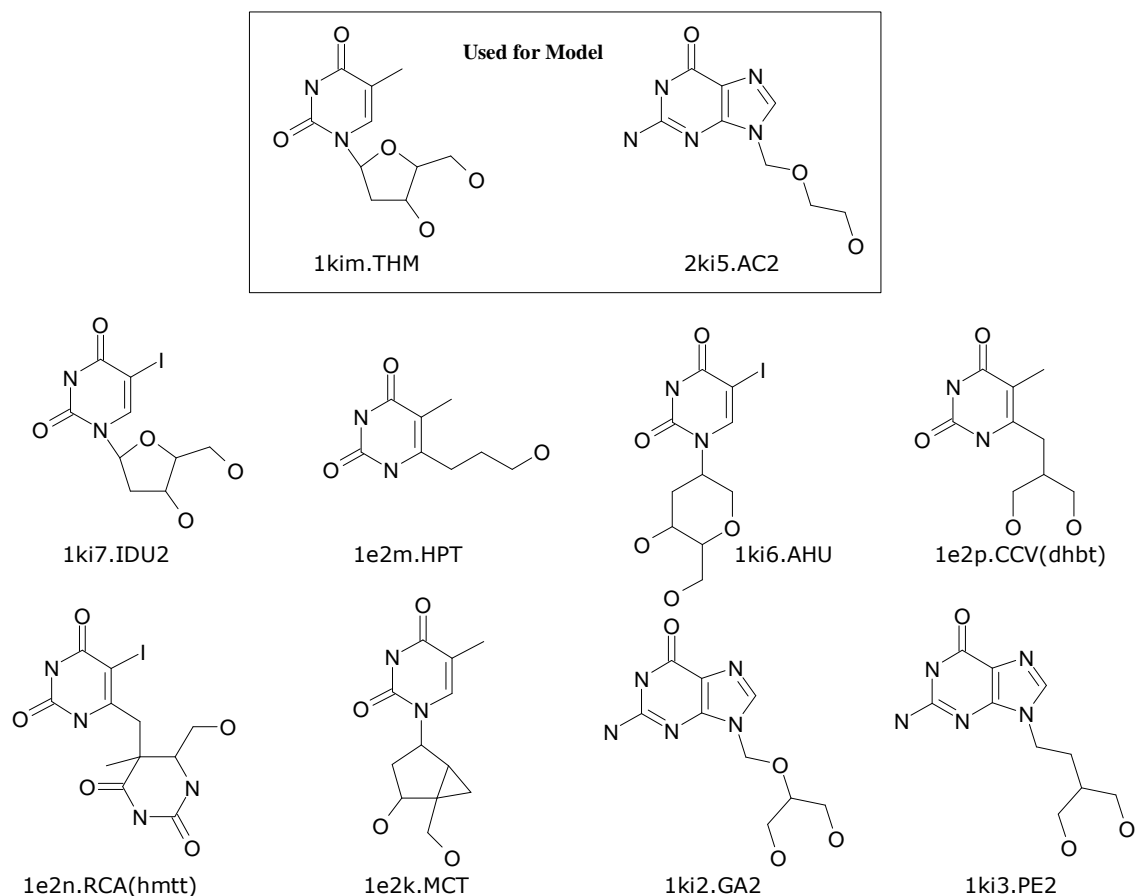


**Figure 2.** The positive compounds in training sets of GPCR and GABA<sub>A</sub>. A and B are serotonin ligands. C (tolterodine), D, and E are muscarinic antagonists. Molecules F–H (bromodiphenhydramine, pyrilamine, and azatadine) are H1 receptor antagonists. Molecules I–K are GABA<sub>A</sub> receptor agonists (diazepam, alprazolam, and zopiclone).

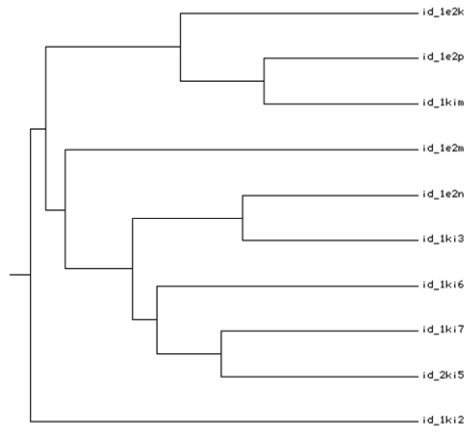


**Figure 3.** The general definition of atom pair descriptors.

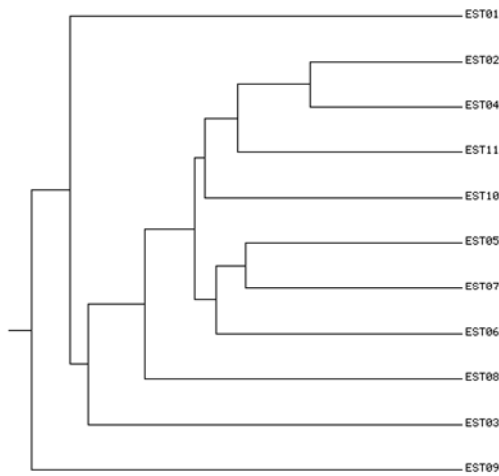




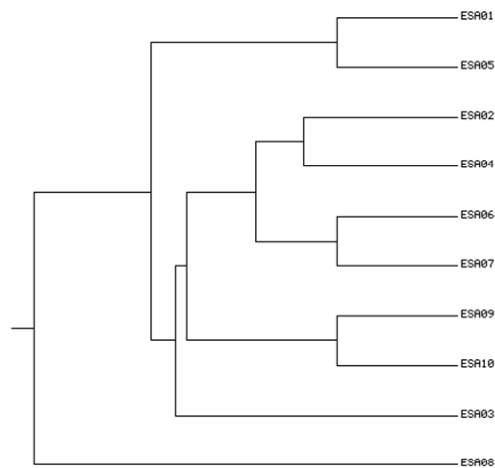
**Figure 4.** 10 thymidine kinase inhibitors were studied for ligand-based screening. These two ligands in the box are the positive cases in the training set and the other eight compounds are used for test. The abbreviations are as follows: 1kim.THM, deoxythymidine; 1ki7.ID2, 5-iododeoxyuridine; 1e2m.HPT, 6-(3-hydroxy-propyl-thymine); 1ki6.AHU, 5-iodoracil anhydrohexitol nucleoside; 1e2k.TMC, (North)-methanocarbathymidine; 1e2n.RCA, (6-[6-hydroxymethy-5-methy-2,4-dioxo-hexa-hydro-pyrimidin-5-yl-methyl]-5-methyl-1H-pyrimidine-2,4-dione; 1e2p.CCV, 6-(3-hydroxy-2-hydroxymethylpropyl)-5-methyl-1H-pyrimidine-2,4-dione; 2ki5.AC2, acyclovir; 1ki2.GA2, ganciclovir; 1ki3.PE2, penciclovir. The ranks of the eight test compounds in the LigSeeSVM predicted model were 1, 2, 3, 4, 5, 6, 7, 11.



**A**

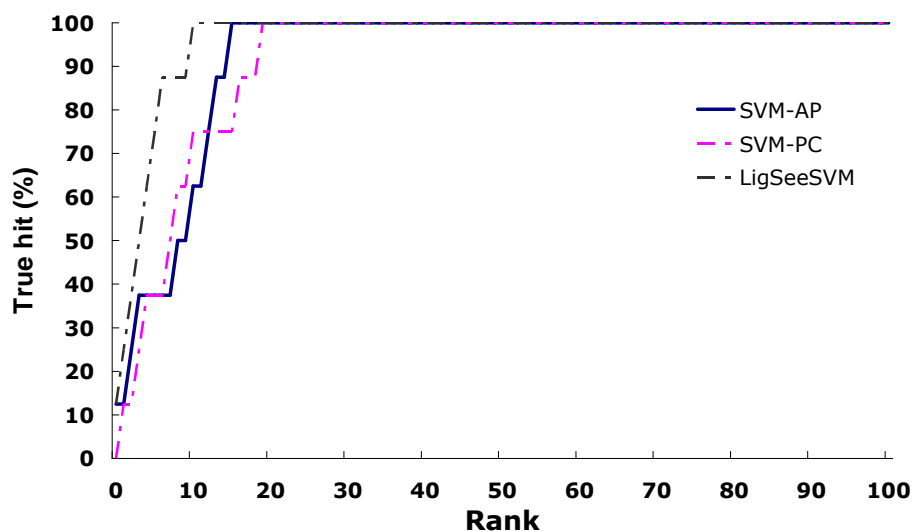


**B**

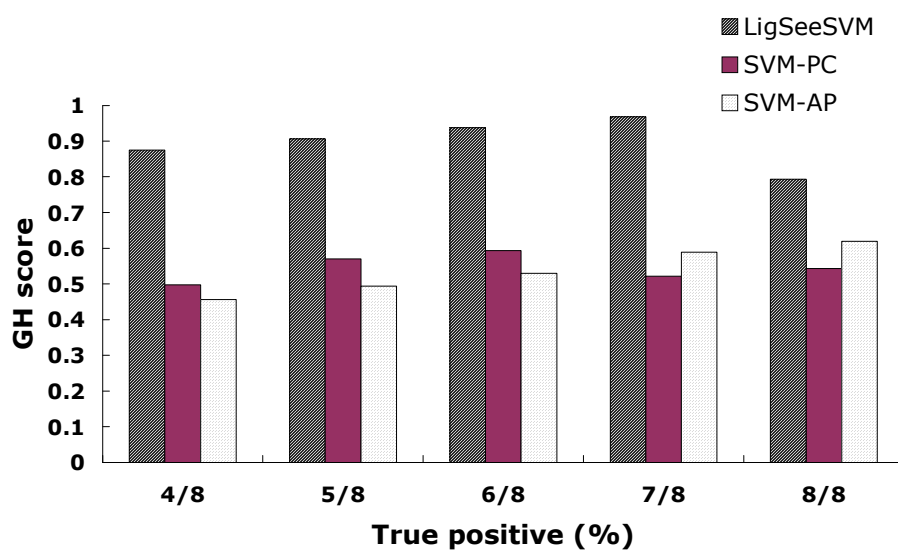


**C**

**Figure 5.** (A) The UPGMA rooted tree of 10 TK inhibitors. (B) The UPGMA rooted tree of 11 ER antagonists. (C) The UPGMA rooted tree of 10 ER agonists.

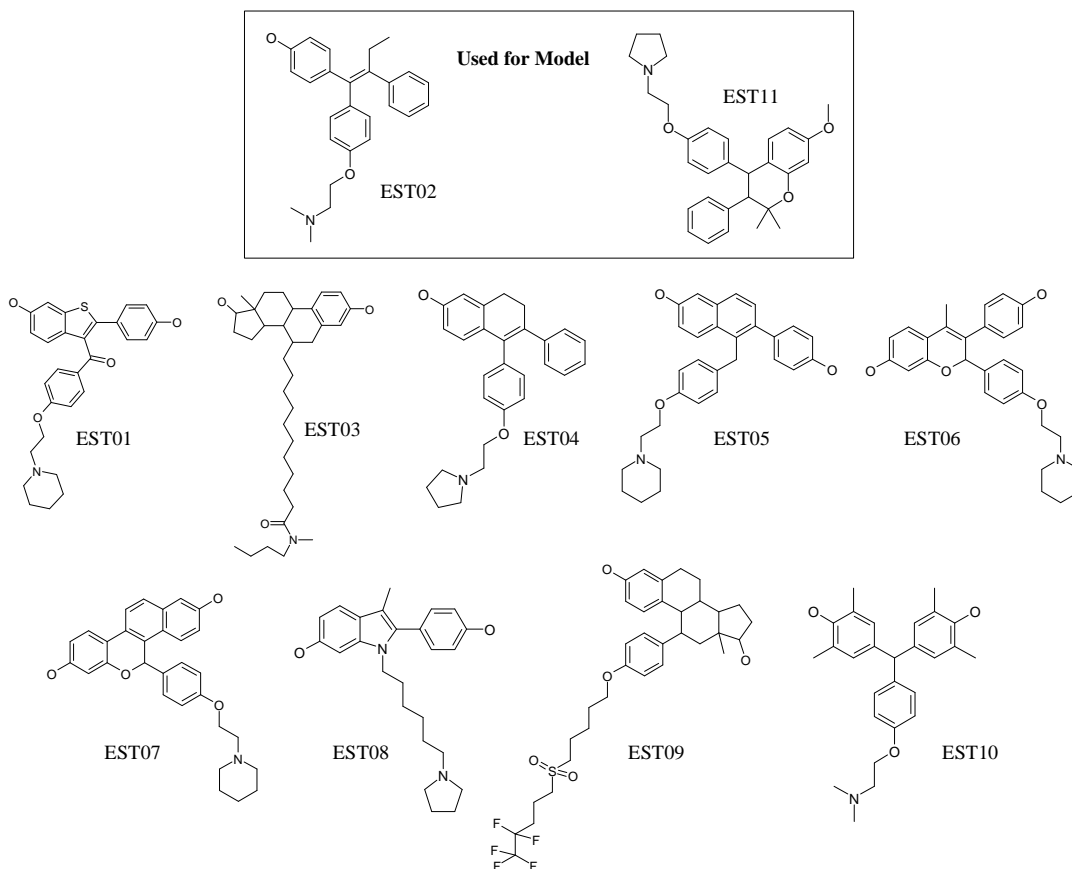


A

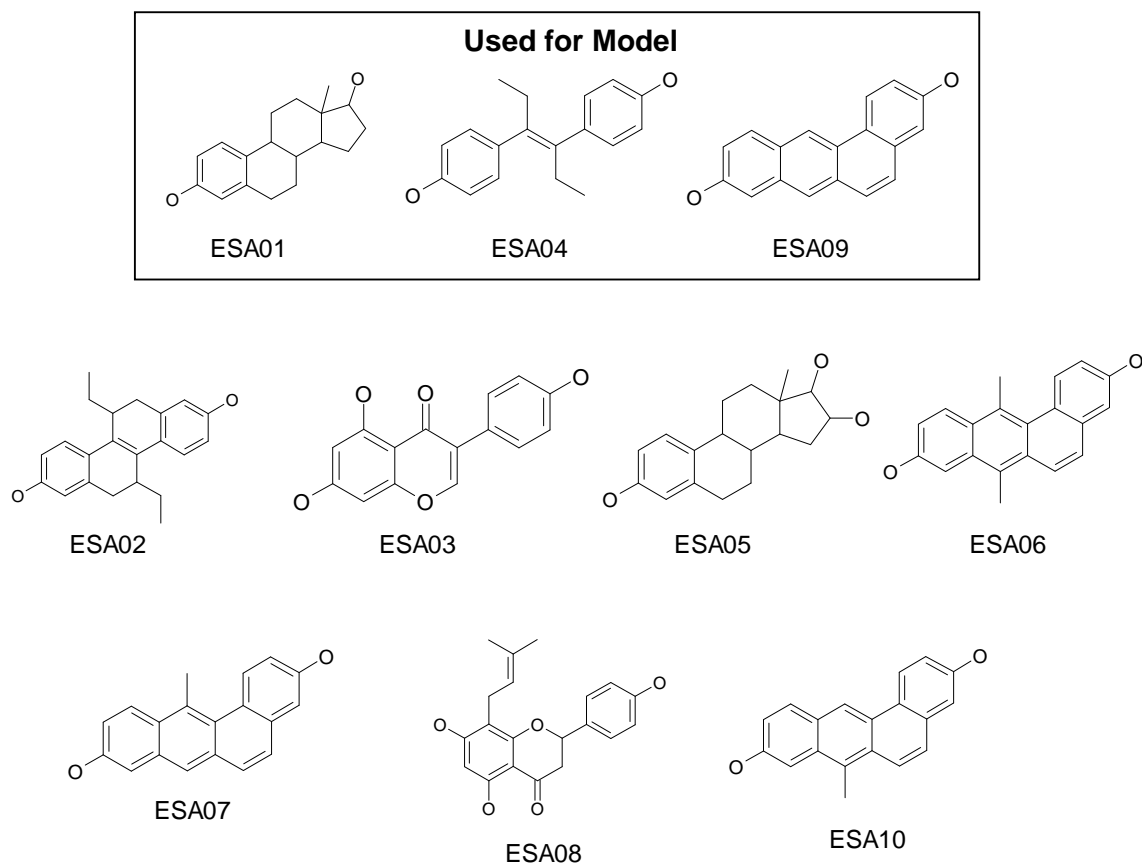


B

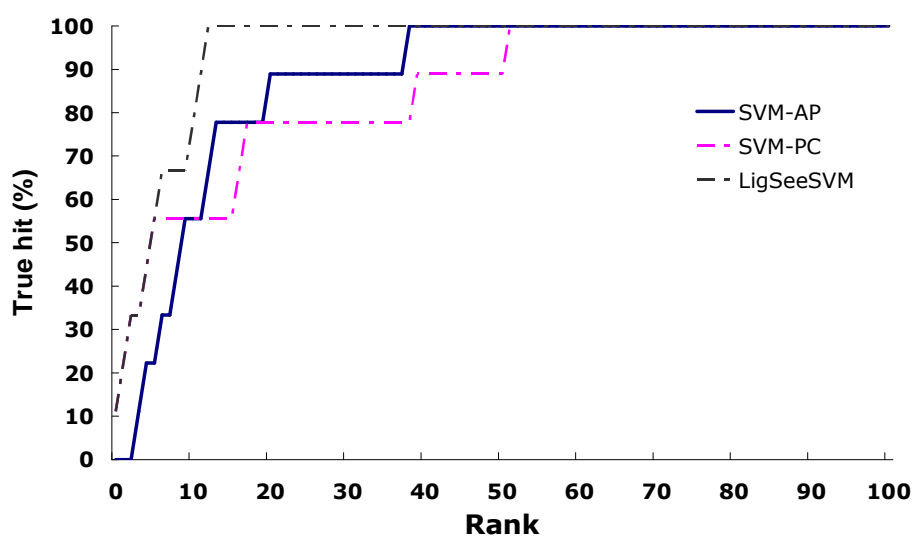
**Figure 6.** The (A) true hits and (B) GH scores of our three SVM models (SVM-AP, SVM-PC, and LigSeeSVM) in screening a TK set with eight positive substrates and 950 negative compounds.



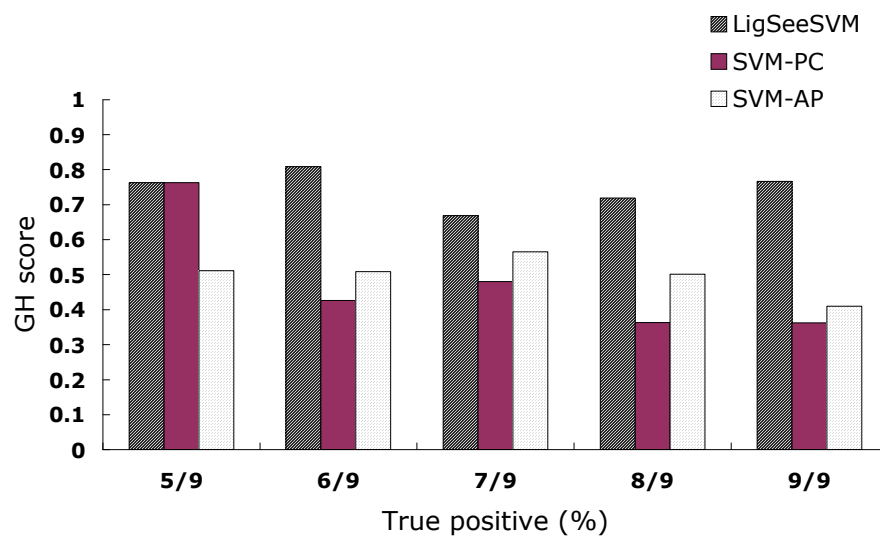
**Figure 7.** 11 estrogen receptor antagonists were studied for ligand-based screening. These two ligands in the box are the positive cases in the training set and the other nine compounds are used for test. The ranks of the nine compounds in the LigSeeSVM predicted model were 1, 2, 3, 5, 6, 7, 11, 12, 15.



**Figure 8.** 10 estrogen receptor agonists were studied for ligand-based screening. These three ligands in the box are the positive cases in the training set and the other seven compounds are used for test. The ranks of the seven compounds in the LigSeeSVM predicted model were 1, 2, 3, 4, 5, 6, 7.



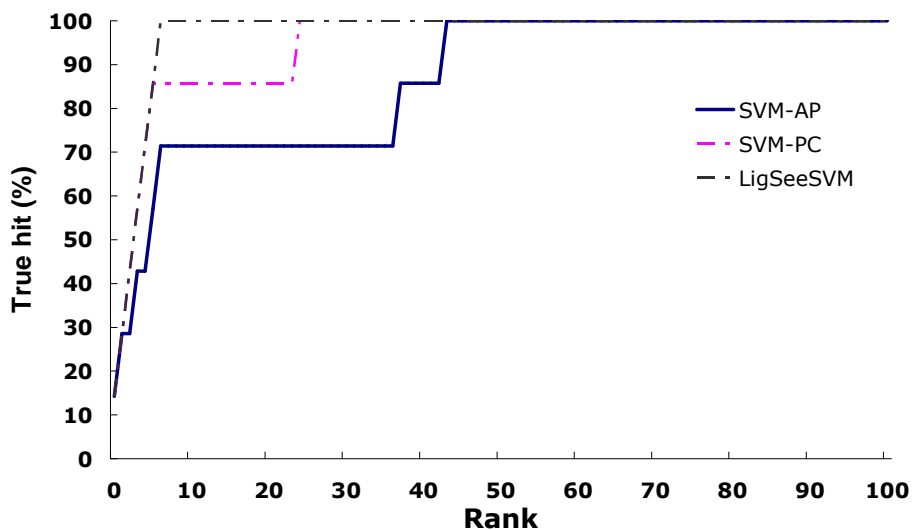
A



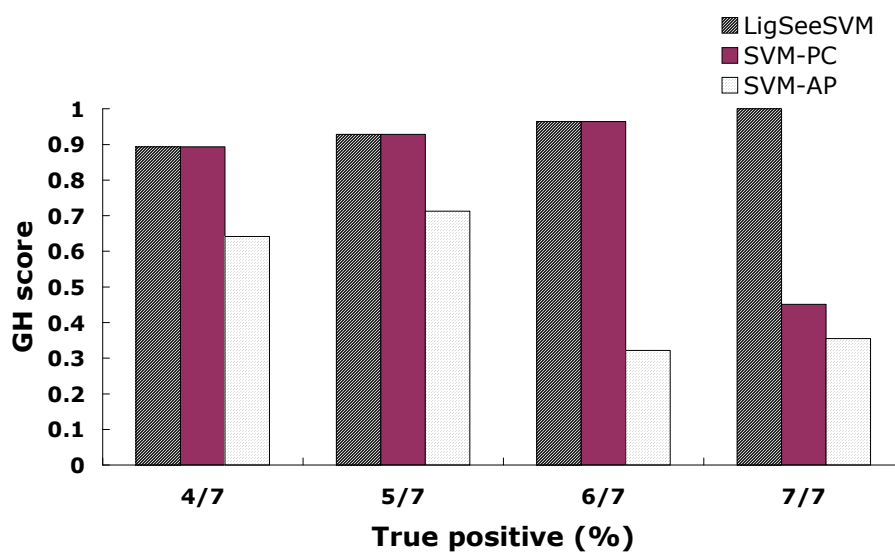
B

**Figure 9.** The (A) true hits and (B) GH scores of our three SVM models (SVM-AP, SVM-PC, and LigSeeSVM) in screening an ER set with nine positive antagonists and 950 negative compounds.



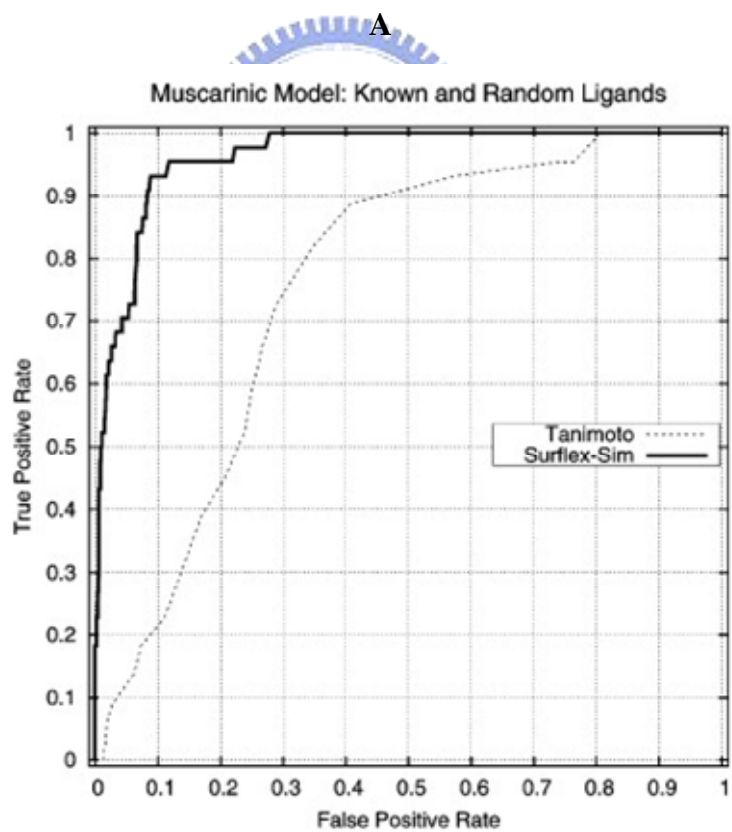
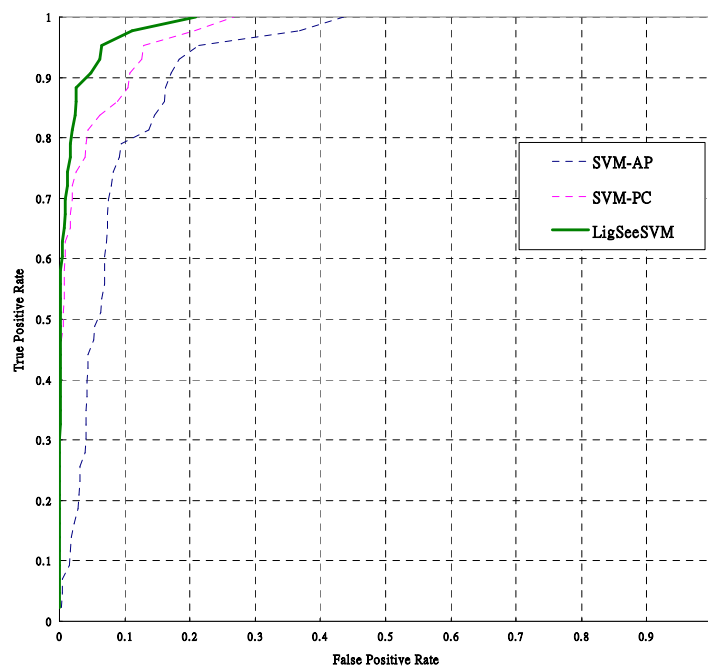


A



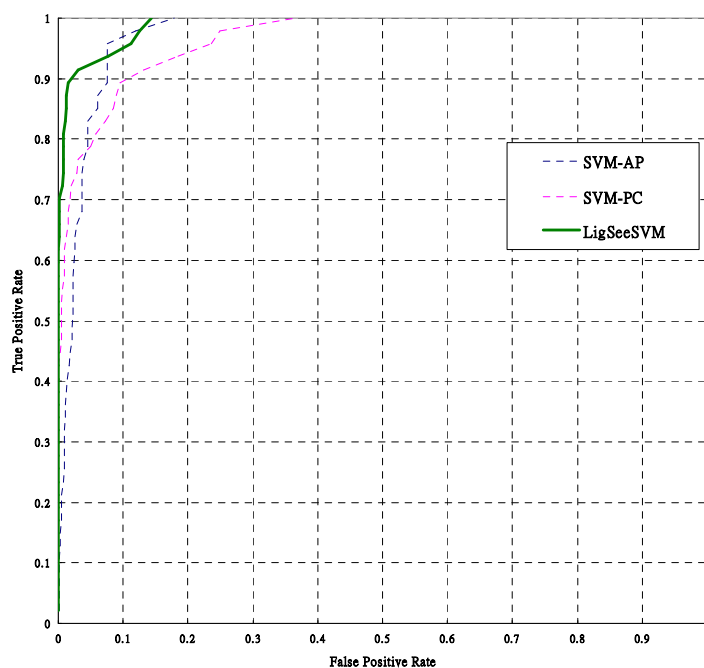
B

**Figure 10.** The (A) true hits and (B) GH scores of our three SVM models (SVM-AP, SVM-PC, and LigSeeSVM) in screening an ER set with seven positive agonists and 950 negative compounds.

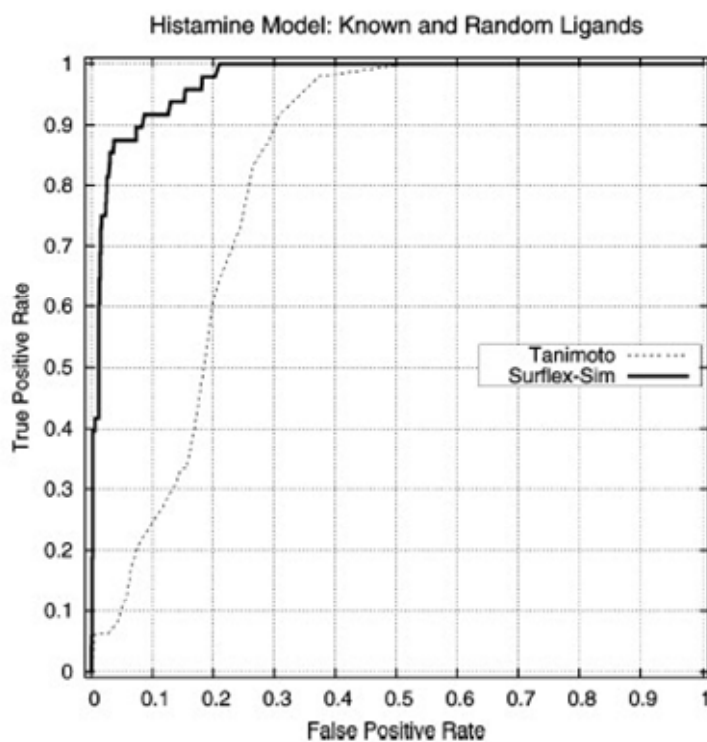


**B**

**Figure 11.** (A) ROC curve of our three SVM models for 43 muscarinic receptor ligands. (B) ROC curve of Surflex-Sim and Tanimoto [3] methods for 43 muscarinic receptor ligands. The LigSeeSVM model performs better than Surflex-Sim.

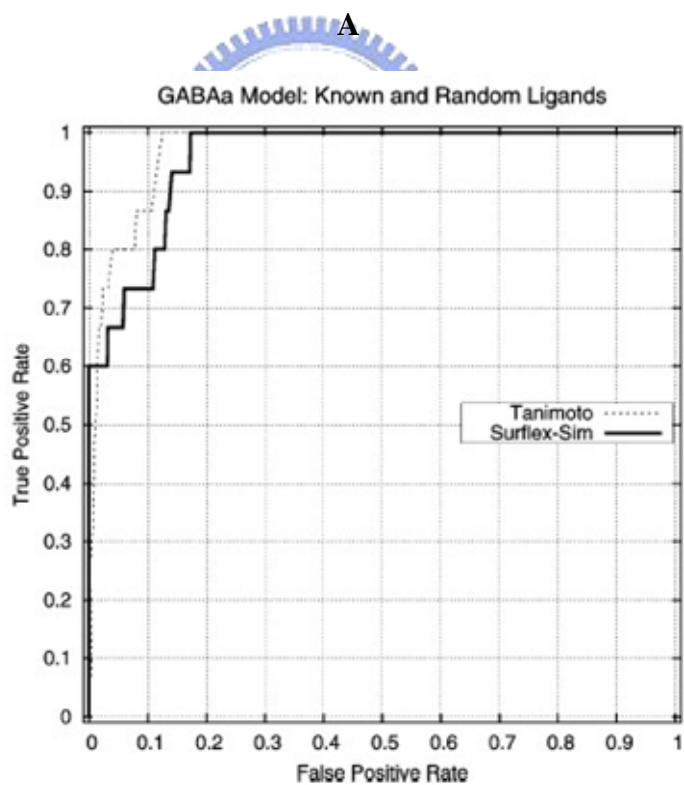
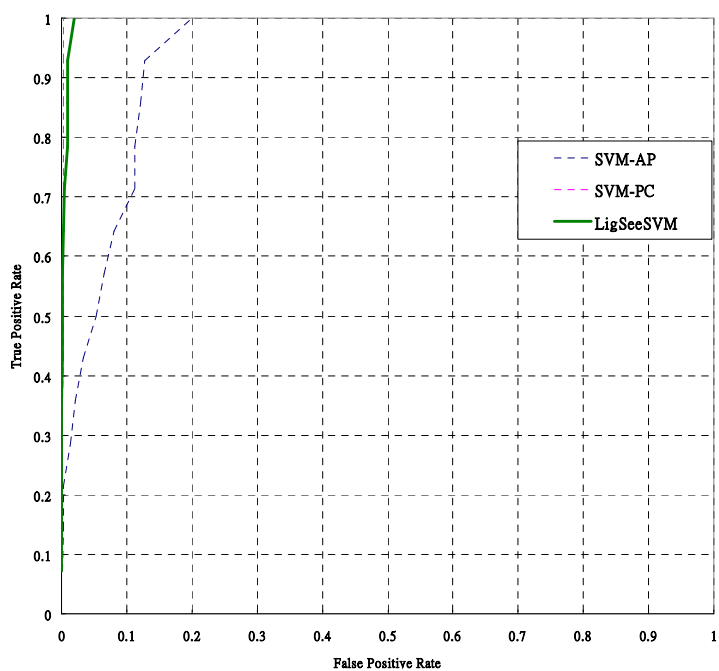


A



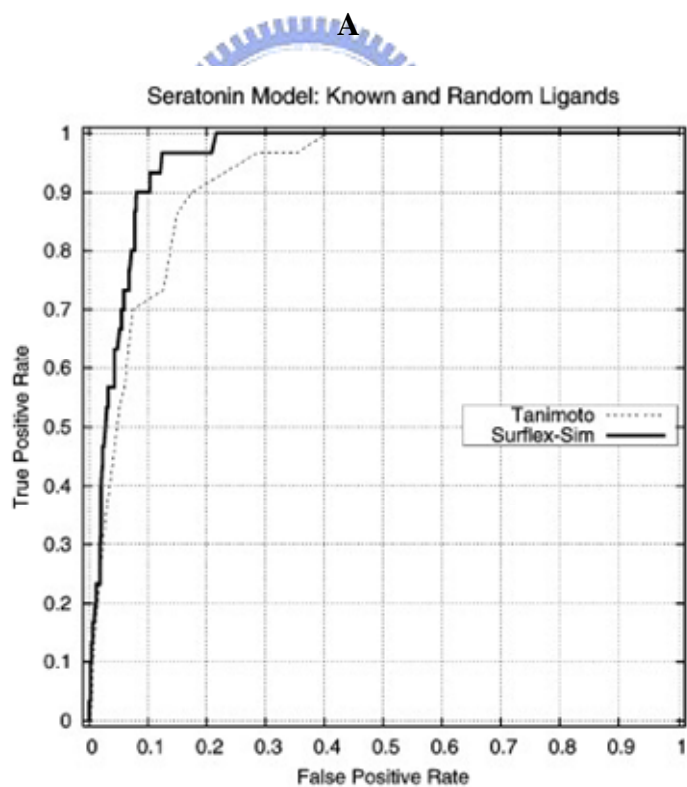
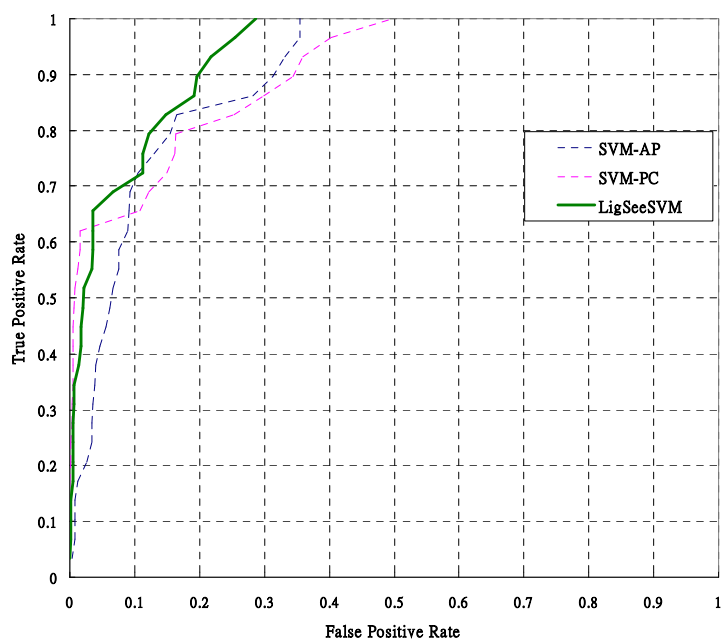
B

**Figure 12.** (A) ROC curve of our three SVM models for 47 histamine receptor ligands. (B) ROC curve of Surfex-Sim and Tanimoto [3] methods for 47 histamine receptor ligands. The LigSeeSVM model performs better than Surfex-Sim.



**B**

**Figure 13.** (A) ROC curve of our three SVM models for 15 GABA<sub>A</sub> receptor ligands. (B) ROC curve of Surfex-Sim and Tanimoto [3] methods for 15 GABA<sub>A</sub> receptor ligands. The LigSeeSVM model performs significantly better than Surfex-Sim.



**B**

**Figure 14.** (A) ROC curve of our three SVM models for 29 serotonin receptor ligands. (B) ROC curve of Surflex-Sim and Tanimoto [3] methods for 29 serotonin receptor ligands.

## References

1. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, 7, 1047-1055.
2. Schneider, G.; Nettekoven, M., Ligand-based combinatorial design of selective purinergic receptor (A<sub>2A</sub>) antagonists using self-organizing maps. *Journal of Combinatorial Chemistry* **2003**, 5, 233-237.
3. Jain, A. N., Ligand-based structural hypotheses for virtual screening. *Journal of Medicinal Chemistry* **2004**, 47, 947-961.
4. Renfrey, S.; Featherstone, J., Structural proteomics. *Nature Reviews. Drug Discovery* **2002**, 1, 175-176.
5. Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z., Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *Journal of Chemical Information and Computer Sciences* **2004**, 44, 1630-1638.
6. Livingstone, D. J., The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Sciences* **2000**, 40, 195-209.
7. Katritzky, A. R.; Gordeeva, E. V., Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *Journal of Chemical Information and Computer Sciences* **1993**, 33, 835-857.
8. Cruciani, G.; Pastor, M.; Guba, W., VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *European Journal of Pharmaceutical Sciences: Official Journal of the European Federation for Pharmaceutical Sciences* **2000**, 11 Suppl 2, S29-S39.
9. Karelson, M.; Lobanov, V. S.; Katritzky, A. R., Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chemical Reviews* **1996**, 96, 1027-1044.
10. Katritzky, A. R.; Tatham, D. B.; Maran, U., Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure-toxicity relationships. *Journal of Chemical Information and Computer Sciences* **2001**, 41, 1162-1176.
11. Zheng, W.; Tropsha, A., Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Sciences* **2000**, 40, 185-194.
12. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A., Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *Journal of Medicinal Chemistry* **2002**, 45, 2811-2823.
13. Accelrys, Cerius<sup>2</sup> 4.8. **2003**.

14. Bock, J. R.; Gough, D. A., A new method to estimate ligand-receptor energetics. *Molecular & Cellular Proteomics: MCP* **2002**, 1, 904-910.
15. Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V., Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of Chemical Information and Computer Sciences* **2003**, 43, 2048-2056.
16. Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences* **2003**, 43, 1882-1889.
17. Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K., Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *Journal of Chemical Information and Computer Sciences* **2003**, 43, 1269-1275.
18. Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A., Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *Journal of Chemical Information and Computer Sciences* **2003**, 43, 2019-2024.
19. Koike, A.; Takagi, T., Prediction of protein-protein interaction sites using support vector machines. *Protein Engineering, Design & Selection: PEDS* **2004**, 17, 165-173.
20. Chang, C. C.; Lin, C. J., Training nu-support vector classifiers: theory and algorithms. *Neural Computation* **2001**, 13, 2119-2147.
21. Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T., An accurate QSPR study of O-H bond dissociation energy in substituted phenols based on support vector machines. *Journal of Chemical Information and Computer Sciences* **2004**, 44, 669-677.
22. Vapnik, V., The Nature of Statistical Learning Theory. *Springer: New York* **1995**.
23. Vapnik, V., Estimation of Dependences Based on Empirical Data. *Springer: Berlin* **1982**.
24. Burges, C. J. C., A tutorial of support vector machines for pattern recognition. <http://svm.research.bell-labs.com/SVMdoc.html> **1998**.
25. Vapnik, V.; Golowich, S.; Smola, A., Support vector method for function approximation, regression, and signal processing. *Advances in Neural Information Processing Systems* **1997**, 9, 281-287.
26. Bissantz, C.; Folkers, G.; Rognan, D., Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **2000**, 43, 4759-4767.
27. Jain, A. N., Surfex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of Medicinal Chemistry* **2003**, 46, 499-511.
28. Sadowski, J.; Gasteiger, J.; Klebe, G., Comparison of automatic three-dimensional model

- builders using 639 x-ray structures. *Journal of Chemical Information and Computer Sciences* **1994**, 34, 1000-1008.
29. van Lipzig, M. M.; ter Laak, A. M.; Jongejan, A.; Vermeulen, N. P.; Wamelink, M.; Geerke, D.; Meerman, J. H., Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *Journal of Medicinal Chemistry* **2004**, 47, 1018-1030.
30. Lin, C. H.; Haadsma-Svensson, S. R.; Lahti, R. A.; McCall, R. B.; Piercey, M. F.; Schreur, P. J.; Von Voigtlander, P. F.; Smith, M. W.; Chidester, C. G., Centrally acting serotonergic and dopaminergic agents. 1. Synthesis and structure-activity relationships of 2,3,3a,4,5,9b-hexahydro-1H-benz[e]indole derivatives. *Journal of Medicinal Chemistry* **1993**, 36, 1053-1068.
31. Lin, C. H.; Haadsma-Svensson, S. R.; Phillips, G.; Lahti, R. A.; McCall, R. B.; Piercey, M. F.; Schreur, P. J.; Von Voigtlander, P. F.; Smith, M. W.; Chidester, C. G., Centrally acting serotonergic and dopaminergic agents. 2. Synthesis and structure-activity relationships of 2,3,3a,4,9,9a-hexahydro-1H-benz[f]indole derivatives. *Journal of Medicinal Chemistry* **1993**, 36, 1069-1083.
32. Jain, A. N.; Harris, N. L.; Park, J. Y., Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5-HT1A receptor. *Journal of Medicinal Chemistry* **1995**, 38, 1295-1308.
33. Nilvebrant, L., Tolterodine and its active 5-hydroxymethyl metabolite: pure muscarinic receptor antagonists. *Pharmacology & Toxicology* **2002**, 90, 260-267.
34. Nordvall, G.; Sundquist, S.; Johansson, G.; Glas, G.; Nilvebrant, L.; Hacksell, U., 3-(2-Benzofuranyl)quinuclidin-2-ene derivatives: novel muscarinic antagonists. *Journal of Medicinal Chemistry* **1996**, 39, 3269-3277.
35. Johansson, G.; Sundquist, S.; Nordvall, G.; Nilsson, B. M.; Brisander, M.; Nilvebrant, L.; Hacksell, U., Antimuscarinic 3-(2-furanyl)quinuclidin-2-ene derivatives: synthesis and structure-activity relationships. *Journal of Medicinal Chemistry* **1997**, 40, 3804-3819.
36. Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G., GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Research* **1998**, 26, 275-279.
37. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R., Atom pairs as molecular features in structure-active studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **1985**, 25, 64-73.
38. Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A., Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *Journal of Medicinal Chemistry* **2004**, 47, 2356-2364.
39. Sneath, P. H. A.; Snokal, R. R., Unweighted pair-group method using arithmetic averages. *Numerical Taxonomy* **1973**, 230-234.



40. Yang, J. M.; Shen, T. W., A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins* **2005**, 59, (2), 205-220.

