THE REVIEW OF VIRTUAL SCREENING TECHNIQUES

Kitsuchart Pasupa

Faculty of Information Technology,

King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

Email: kitsuchart@it.kmitl.ac.th

ABSTRACT

Effective treatments extend lives in the world, and significant efforts are in place to expand the use of life-saving medications in the developing world. This paper gives an overview of drug discovery process and emphasises in the area of virtual screening. Because machine learning is fast becoming a popular mechanism to support activity recognition in drug discovery process and other real-world applications, hence, this review emphasises on machine learning methods used in virtual screening. This includes linear and kernel discriminant analysis, neural network, decision tree, graph kernel machines, and support vector machine.

Index Terms – drug discovery; chemoinformatics; virtual screening; machine learning

1. Introduction

The greatest accomplishments in mankind, arguably, so far has been in the field of medicine and fighting disease. One of the public health's greatest achievements is the eradication of smallpox which was discovered by Edward Jenner in the late 18th century [1]. He made use of cow-pox virus to build immunity against the deadly scourge of smallpox [2]. He invented the term "vaccination" for his treatment from the Latin word "vacca"– a cow. The vaccination was adopted by Louis Pasteur for immunisation against any diseases later on [2]. The other greatest achievement is a discovery of Penicillin by Alexander Flemming in 1929 [3]. His observation that a fungal contaminant had wiped out a bacterial colony resulted in Penicillin, which has subsequently saved a millions of lives and limbs: more people died in battle than from infection in World War II, for the first time in any war [4].

People have used drugs for as long as they have tried to relieve the pain. Drugs are supposed to cure disease or improve health but they are also able to damage it, e.g. cocaine, amphetamines, heroin. Drugs are molecules that work by interacting with others in the body [5]. The human body consists of around a hundred trillion cells. Cells extract energy from food and oxygen to grow and divide, and also to send signals to other cells within the body by the use of protein. Drugs are able to influence the

function of cells by affecting one or more different proteins effecting growth, division, metabolism and routine signalling. A cell controls the function of its proteins by other molecules. The signaling molecules can be other proteins, protein derived molecules, etc. These signaling molecules and protein targets have unique shape which allows them to fit together. A drug can mimic a signalling molecule if its shape is similar enough to one. Once the drug fits into the drug's target, it can either activate or block its target.

[5] gives a concrete example of how drugs treat a failing heart. The heart beats very fast, during its failure, to compensate for its inefficient pumping. Thus the heart rate needs to be slowed down in order to reduce damage. An important signalling molecule, "adrenaline"is able to bind to a protein found in heart cells, beta-adrenergic receptor. When they are bind together, the heart will beat faster. Hence, if a drug has a similar enough shape to adrenaline, it would fit into a receptor. If it has a very similar shape to adrenaline, it would activate the receptor just like adrenaline does. However, if its shape is similar and yet somewhat different than adrenaline, it would fit into receptor and not activate it. This leads to a slowing of the heart rate.

Drug discovery is a very complex task and time consuming with a recent estimate by the Tufts Centre for Drug Development [6] suggesting that it takes 12–15 years to develop a drug at a cost of 803 million dollars.

The process of developing a drug starts from sample collection and selection of targets for drugs. High-throughput technologies e.g. high-throughput screening and combinatorial chemistry are then used (see section 3 for more details), followed by lead optimisation. The purpose of lead optimisation is to optimise the molecules or compounds which demonstrate the potential to be transformed into drugs, retaining only a small number for the next stages. It can be done by e.g. *in silico* (on computer) modelling [7], X-ray crystallography [8]. A set of selected compounds is then tested both *in vitro* (within the glass) and *in vivo* (within the living) in the preclinical testing stage. The testing is usually conducted using animals e.g. mice. Mice are widely used in research because humans share approximately 99% of genes with mice [9]. At this stage, a set of selected compounds is reduced to only 5–10 compounds that are relevant enough to test on humans in the clinical testing stage before a drug is launched on to the market. Figure 1 shows the process of developing a drug. The diagram is developed from [10] and [11].

2. CHEMOINFORMATICS

Chemoinformatics is the use of computer and informational techniques to tackle chemical problems which emphasise the manipulation of chemical structural information [7]. Because chemoinformatics is quite new, there is no universal agreement on the correct spelling. It is also known as *cheminformatics*, *chemiinformatics*, and *chemical informatics*.

The term was first defined by [12] as "the mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimisation" [12]. In fact, chemoinformatics is simply a new name, it is not a new discipline [13].

Many techniques used in chemoinformatics have been well-documented and reviewed in many text books i.e. [7, 14, 15]. They are used in pharmaceutical companies in the process of drug discovery such as virtual screening and quantitative structure-activity relationship (QSAR).

An important characteristic of the techniques in chemoinformatics is that it must be applicable to a huge amount of data and information (number of molecules) [7]. The data can only be processed and analysed by computer methods which also depend on computational power.

This paper emphasises virtual screening techniques which are explained in the following section.

3. VIRTUAL SCREENING

In the twentieth century mankind has obtained the ability to discover highly active, yet small, organic molecules which are used for treatment purposes. Combining synthetic organic chemistry and information from clinical chemistry enables us to develop the powerful medicines available today. However, medicinal chemists have always struggled with the selection of which compounds to synthesise : a chemist has to choose the compounds to be synthesised from among millions of possible molecules.

New technologies, the so-called "combinatorial synthesis"and "high-throughput screening" have been introduced because, in the past, a few hundreds of compounds could be synthesized by one chemist in a year [16]. Potential drug compounds were normally obtained from natural products or from scientific literature on known compounds. Combinatorial synthesis offers a much broader range of possibilities. It leads to an increase in the number of compounds in the pharmaceutical company screening libraries into the millions. Analysis of a specific biological activity for an entire screening library can be done by highthroughput screening in a matter of days. In order to control costs, time and waste, the computational chemist is encouraged to develop some kind of computer programme capable of automatically evaluating very large libraries of compounds and integrate it into the drug discovery process. This is called "virtual screening" (VS).

VS is a set of computational methods or *in silico* analogues of biological screening. The aim

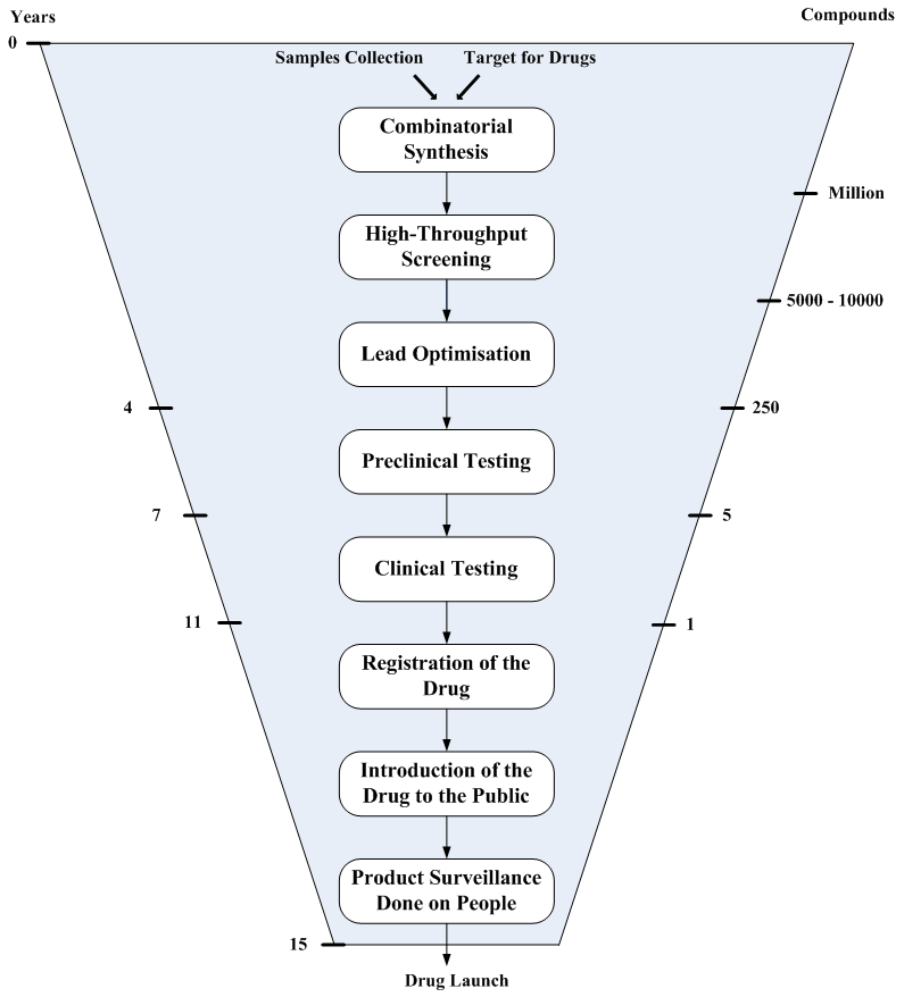

Figure1. Drug discovery process.

of VS is to score, rank and/or filter a set of chemical structures using one or more computational procedures in order to ensure those molecules with the largest prior probabilities of activity are assayed first in a "lead discovery programme"[16, 15, 7]. [17] grouped VS methods into four main classes based on the amount of structural and bioactivity data available, as follows:

1) If just a single active molecule is available, then similarity searching can be used.
2) If several active molecules are known, it is possible to define what is known as a common 3D pharmacophore leading to a 3D database search.
3) Machine learning methods should be used for VS only if it is not possible to identify a common pharmacophore and there is a sufficient number of active and inactive molecules available. They can be used to derive a structureactivity relationship from the known active molecules for use in

predicting biological activity. An example of machine learning for virtual screening is shown in figure 2.

4) A docking study can be employed if the 3D structure of the biological target is known. It involves the prediction of the binding mode of individual molecules. It aims to identify the orientation that is closest in geometry to the observed structure.

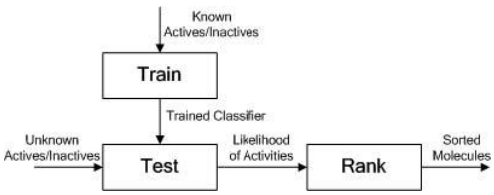

Figure 2. An example of machine learning for virtual screening.

4. TECHNIQUES IN VIRTUAL SCREENING

VS can be divided into two distinct categories: ligand-based VS and structure-based VS [7]. A ligand is a molecule that is able to bind to and form a complex with a biomolecule to serve a biological purpose. Ligand-based VS involves using information available from a single or set of compounds which have been identified as potential leads. A lead compound is a compound that exhibits pharmacological properties which suggest its value as a starting point for drug development. Ligand-based VS is conducted by identifying molecules that share some similarity or properties with the single/multiple active molecules. It aims to score database molecules based on their overall shape similarity to query molecules. Examples of ligand-based VS are substructure/similarity searching, pharmacophore-based designs, and machine learning techniques, which correspond to the first, second, and third class of VS methods grouped by [17] respectively (see section 3). An example workflow for ligand-based VS is shown in figure 3. Structure-based VS can be implemented if the 3D shape binding of the biological target is known. An example of structure-based VS is docking – the fourth class of VS [17].

The scope of this review focuses on techniques which have been developed for VS – in particular machine learning techniques. The section is organised as follows. Sections 4.1 and 4.2 presents an overview of similarity-based VS and pharmacophore-based designs, respectively. Section 4.3 reviews machine learning techniques used in VS followed by the docking method in section 4.4.

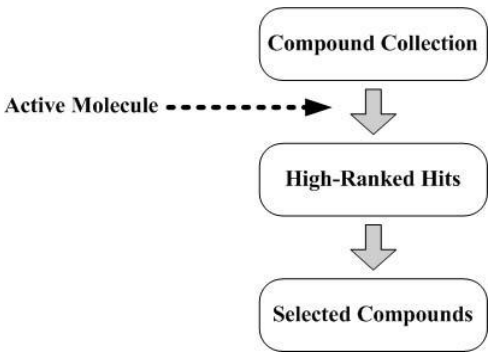

Figure 3. An example work flow for ligand-based VS.

4.1. Similarity-Based Virtual Screening

4.1.1. Similarity Searching

Similarity searching is used for finding those compounds which are most similar to a query

compound in a database. This involves comparing the query compound with every compound in the database in turn and returns a ranked list of all the compounds that are judged to be similar to the query. Similarity searching in chemical databases was first introduced in the mid-1980s [18, 19]. Many methods have been developed such as RASCAL [20], and LINGO [21], involving various descriptors and similarity coefficients. The rationale for similarity searching is the "similar property principal" which states that structurally similar molecules will exhibit similar properties and biological activity [22]. Recently, [23] conducted an experiment on the Daylight fingerprint [24]. Their experiment shows that compounds with similarity values higher than a threshold (0.85 using Tanimoto coefficient) for shared biological activity have only 30% chance of shared biological activity. This is lower than the one that was later accepted by computational chemists [23], emphasising that similarity searching is probabilistic in nature, hence, perfect results cannot be expected.

The most commonly used similarity method is based on 2D fingerprints and there are numerous studies and reviews of similarity coefficients [25, 26]. Similarity coefficients can be classified into three major classes namely: association coefficients, correlation coefficients, and distance coefficients [26]. [27] investigated 36 similarity coefficients and found that the so-called Tanimoto coefficient performed best in similarity search.

Apart from 2D fingerprint-based methods, similarity matching is also used for graphical descriptors. One algorithm that can compare objects represented as a graph is the Maximum Common Subgraph (MCS). The MCS is the largest set of atoms and bonds in common between two structures. The problem of identification of the MCS in two graphs is NP-Complete[1] [7]. A number of exact and inexact methods have been introduced for the MCS problem [28]. [29] first applied MCS matching to database searching by using a two-stage approach: determine an upper-bound on the size of the MCS by using fragment-based search, and then use MCS calculation only on those molecules above a given threshold [29]. Recently, a new algorithm the so called "RASCAL"(Rapid Similarity CALculation) which is based on exact graph matching has been introduced [20]. This algorithm is able to perform tens of thousands of comparisons in a very short time.

[30] evaluated both graph-based and fingerprint-based measures of structural similarity. The results show that, in VS, there is no statistically significant difference in the number of active molecules retrieved by graph-based and fingerprint-based approaches.

However, graph-based approaches provide an effective complement to the fingerprint-based approach. They suggest that there is a way in which these two approaches could be combined: first, generate the initial output

[1] The most difficult problem in NP (non-deterministic polynomial time). They are the smallest subclass of NP which remains output P.

ranking using similarity matching, then use RASCAL to visualise the similarities between the target structure and the nearest neighbours from that ranking. The results show that only 4% of the fused hit-lists contained fewer actives than either of the original hit-lists.

Another type of similarity search system is text-based molecular description. Recently, [21] introduced a new algorithm into QSAR[2] model, LINGO, based on the fragmentation of SMILES strings into overlapping substrings of a defined size. SMILES is a way to represent a molecule through the use of a linear notation. SMILES stands for Simplified Molecular Input Line Entry Specification. SMILES strings are the most compact text-based molecular representations. They contain most of the information that is needed for computing all kinds of molecular structures. The following year, [31] applied LINGO to VS tasks by integrating LINGO into a pseudo-evolutionary algorithm. The results show nearly 10 times better performance than a random search.

4.1.2. Data Fusion

Data fusion is defined as the use of techniques that combine data from multiple sources in order to improve on individual results. It was first applied to the similarity searching of chemical compounds by [32]. The fusion of similarity measures offers a more consistent level of searching performance than just a single measure. There are many different rules by which two or more similarity measures can be combined, e.g. SUM, MAX, and MIN. These are based on those identified by [33]. The SUM method scores each molecule by using an average rank position from each similarity measure. While the MAX and MIN methods score each molecule by using the maximum and minimum rank position obtained from the different measures.

A set of 22 similarity coefficients was divided into 13 groups by using the Mojena stopping rule [34], while in a previous study, it was divided into 11 groups [35]. The 13 groups fusion was tested on the retrieval of bioactive molecules from the NCI AIDS database[3], the IDAlert database[4], and the MDL Drug Data Report (MDDR) database [36]. The results show that applying data fusion to the similarity coefficients improves search performance with little extra computational cost. The optimum numbers of coefficients to use in data fusion has been found, in practice, to be between two and four, with improvement diminishing at five or more. However, there is no single combination which produces a consistently high performance

[2] They represent an attempt to correlate structural or property descriptors of compounds with specific biological activity targets. There are many types of possible model e.g. mathematical and statistical.

[3] The NCI AIDS database is available from NCI/NIH Development Therapeutics Programme at URL http://dtp.nci.nih.gov. It contains information on selected compounds found to be active in the National Cancer Institute's AIDS antiviral screen.

[4] The IDAlert database is available from Current Drugs Limited at URL http://www.current-drugs.com. It contains 11,607 biological activity structures which have been reported in the literature during 1992-96.

across all types of activity classes[5] in the databases.

Data fusion is also the basis for the consensus scoring approach used in protein-ligand docking [7]. Consensus scoring combines multiple scoring functions and leads to higher hit-rates in VS. For consensus scoring there are three strategies to rank molecules according to the predicted results, namely rank-by-number, rank-by-rank, and rank-by-vote [37]. In rank-by-number all molecules are ranked according to the average predicted values given by all the scoring functions. All the molecules are ranked according to the average ranks predicted by all scoring functions in rank-by-rank. If a molecule is predicted to be on the top, e.g. 5%, for each scoring function, then it is given a vote. The final score of each molecule is the sum of votes received from all scoring functions. This is called rank-by-vote.

It was found that consensus scoring can dramatically reduce false positive predictions [38]. Moreover, only three or four scoring functions are sufficient for consensus scoring [37]. Among the three strategies described above, rank-by-number and rank-by-rank work more effectively than rank-by-vote. This has been shown in an extensive study in [37].

[39] combined consensus scoring with the Näive Bayes classifier in order to improve enrichment[6] of high-throughput docking results. In this study, rank-by-median was introduced and found to be more effective than rank-by-mean and rank-by-vote. In rank-by-median all molecules are ranked according to the median of predicted values given by all the scoring functions. Using the rank-by-median with Näive Bayes classification is robust enough to ensure maximum enrichment [39].

4.2. Pharmacophore-Based Designs

In the early 1900s, a pharmacophore was first defined by Paul Ehrlich as “a molecular framework that carries (phoros) the essential features responsible for a drug’s (pharmacon) biological activity” [40]. The definition was updated by Peter Gund in 1977 as “a set of structural features in a molecule that is recognised at a receptor site and is responsible for that molecule’s biological activity” [40]. Recently an International Union of Pure and Applied Chemistry working party elaborated the definition as “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response” [41]. Pharmacophores are used to define the essential features (with 3D arrangement) of one or more molecules with the same biological activity in computational chemistry. The initial

[5] An activity class is defined as a specific biological activity target.

[6] Enrichment is defined as the curve depicting number of actives retrieved experimentally versus number of samples retrieved by random search.

studies focused on rigid 3D structures within each molecule in a database being represented by a single and low-energy conformation. A pharmacophore search is likely to miss large numbers of matching molecules that can adopt a conformation containing the query pattern but that are represented in the database by a low-energy conformation that does not contain this pattern. This is because it takes no account of the flexibility that characterises many molecules. Hence, some approaches have been introduced in [42] to flexible 3D searching.

Many pharmacophore algorithms have been developed, e.g. DISCO (DIStance Comparison) [43] makes use of graph matching algorithm – MCS algorithm, and GASP (Genetic Algorithm Superimposition Program) [44] uses a genetic algorithm to identify possible pharmacophore models from small sets of active compounds (typically 2–5 compounds).

The use of pharmacophores is reflected in the large amount of literature. A number of textbooks dedicated to pharmacophores have been published e.g. [40, 45]

4.3. Machine Learning Techniques

Machine learning involves the design and development of algorithms and techniques in order to allow computers to learn and understand data. Machine learning research focuses on extracting the relationship from available data by computational and statistical methods. There are many reasons why engineers need to identify or model complex relationships e.g. understanding the world, predicting the future, classification, decision support and control. A machine learns when it changes its structure or data based on its input, hence, broadly say, performance of the machine is expected to be improved, assuming that data are plentiful.

4.3.1. Substructural Analysis

Substructural analysis is the first machine learning method used in chemoinformatics [46]. It is a class of QSAR techniques which assume that a defined molecular fragment gives a constant contribution to an activity. It aims to assign a weight for each substructural fragment which reflects its possibility of being active or inactive. A score for each unknown compound can be calculated from the sum of the weights of all of the fragments contained within a molecule. Unknown compounds are then ranked, based on the calculated scores, in decreasing probability of activity.

Many fragment weighting schemes have been proposed and compared [47]. Substructural analysis was neglected for many years but is subject to renewed interest [48, 49]. One of the reasons is that the weighting schemes used in substructural analysis are very similar to those used in Näive Bayesian classifiers one of the most widely use machine learning techniques [50].

4.3.2. Linear Discriminant Analysis

Here, linear discriminant analysis (LDA) aims to separate molecules into defined classes. The simplest type of LDA is the binary classification problem. It finds the linear combination of

features and draws the hyperplane which best separates the data. It was first applied to a set of biologically active molecules by [51]. Their study involves a relatively small dataset containing only 20 compounds with general structure of the aminotetralins and aminoindans.

Recently, [52] performed VS based on LDA to identify new lead trypanothione reductase inhibitor compounds. The database is represented by 2D and 3D descriptors which are then combined together in order to capture complementary information. They used 58 compounds for training set and 422,367 compounds for test set. The experiment shows good results with 91.38% and 88.63% for training and test set, respectively. More applications of LDA to VS are described in [53] and [54].

4.3.3. Neural Networks

Neural networks (NNs) are a type of artificial intelligence that attempt to mimic the way a human brain works. They work by creating connections between artificial neurons which are arranged in layers and associated with weights. The weights are initially set to random values. The NN must first be trained in order to adjust the weights by using a known set of inputs and corresponding set of outputs[7].

NNs are trained repeatedly until they can obtain the chosen performance criterion for the training set. Once a NN has been trained and validated, it can be used to predict activity for unknown molecules. NNs are a common working tool for problem solving in computational chemistry [55, 15, 56, 57, 58]. Feed-forward networks and the Kohonen network are the two most commonly used NN architectures in chemistry [7]. Feed-forward networks use a supervised learning method which derives the model by using the values of the dependent variables, while the Kohonen network is unsupervised.

One problem with the use of NNs is overtraining. They can give outstanding results on the training set, however, when used with unknown data, they can give poor predictions because the training data were effectively memorised by the network. In order to address this problem, *M*-fold cross-validation is used.

4.3.4. Decision Trees

Decision trees consist of a set of rules which can associate specific molecular features with an activity or property of interest. A decision tree represents a Boolean function and is described as a tree-like structure. Each node corresponds to a specific rule. Once decision trees have been trained from a training set, an unknown sample can be classified by starting at the root node then following the edge appropriate to the rule. This is repeated until a terminal node is reached. This approach allows easy determination of the most relevant chemical features to the target biological property [59]. Various methods are available to construct decision trees such as ID3 [60], C4.5 [61, 62], and C5.0 [61].

Applications of decision trees to drug discovery are addressed in [7]. Recently, [63]

[7] The outputs are needed for supervised learning.

compared seven different classification methods on the selection of active molecules for five different target protein problems. The experiment shows that decision trees slightly outperform only Näive Bayesian classifier but other methods e.g. support vector machines, NNs achieve better accuracy than decision trees.

4.3.5. Support Vector Machine

The support vector machine (SVM) is a more recent example of a powerful machine learning algorithm for solving chemical classification problems [64, 65, 66, 67]. It is a supervised learning algorithm first introduced by [68]. It uses a linear decision boundary to discriminate between two classes. The extension of the SVM to multi-class classification, regression, and probability density estimation also exists.

A training data set which consists of two classes is represented by an *m*-length binary fingerprint. The SVM works by implicitly mapping the training data set into an *N*-dimensional feature space ($N > m$) by the so-called kernel function. Then it identifies a large-margin hyperplane which separates the two classes of data. Once the machine has been trained and validated, it can be used to discriminate unknown data.

The SVM in conjunction with Fisher kernel was successfully applied to protein classification by [69]. They begin by training a generative hidden Markov model[8] (HMM) in order to model a given protein family. This maps all protein sequences to points in a Euclidean feature space of fixed dimension. The SVM, in conjunction with the Fisher kernel is then used for protein classification.

In a study by [66], an SVM is compared with NN on a large benchmark dataset containing atomic descriptors of compounds that are drugs or nondrugs. They found that SVM (20% error rate) is only slightly better than a simple NN (20.75% error rate).

[65] suggest that a careful model selection procedure can improve dramatically upon existing results. They are able to improve on the result of [66] and reduce the error rate to 18.1% by taking a careful approach to model selection. In the study of [65], the SVM is compared with other modern classifiers. After careful model selection, a blind test is used. By testing the selected learning machines on unseen data, the SVM with polynomial kernel of degree $p = 11$ and with RBF kernel with $\sigma = 5$ yielded error rates of 7.1% and 6.9%, respectively.

A serious problem with the SVM in conjunction with the Fisher kernel is its computational expense. The cost of computing each kernel entry over strings is $O(m^2)$ in the length of the input sequence. [67] introduced the SVM in conjunction with a string kernel (spectrum kernel) for the protein classification problem which allows linear time classification with complexity $O(km)$ to compute each kernel entry, where k is a length of spectrum of the input sequence. They suggest that the SVM in conjunction with a string-based kernel could offer a simple, effective and computationally

[8] HMM is a way of extracting features from protein sequences.

efficient alternative to other methods of protein classification.

[70] applied SVM to a *heterogeneous* (sometimes called *diverse*) set of active compounds. Heterogeneous defines diversity among substructures of molecules in a database. In other words, molecules in database are not alike. Their experiments show that the SVM is more effective than binary fingerprint-based ranking methods e.g. binary kernel discrimination. Moreover, they also improved the performance of VS by combining the results of the SVM and binary kernel discrimination using data fusion techniques. SVM calculated scores for 25,300 test molecules in 15 seconds while binary kernel discrimination spent approximately two minutes.

Probabilistic SVM [71] in conjunction with the ECFP_4 fingerprint [72] was applied to construct a drug-likeness filter for molecules [73]. They compared their results with previous published results by [66]. They are able to reduce the error rate to 7.27%.

SVM now plays a role in VS tasks and there is a growing literature on SVMs in VS e.g. [74, 75, 76, 77, 78, 79, 62].

4.3.6. (Binary/Continuous) Kernel Discrimination

Binary kernel discrimination (BKD) is a machine learning classification tool that has been successfully applied to VS tasks. [80] introduced a multivariate binomial distribution into the methods of kernel density estimation (KDE) [81]. BKD was first applied in chemoinformatics by [82]. KDE is a nonparametric method and so makes no assumptions about the frequency distributions of the variables being assessed [83]. The aim of KDE is to estimate the true probability density function of a given sampled data. In BKD, KDE is used to estimate the distribution of a sample of molecules from a training set in order to describe the physical or structural properties of molecules in some multidimensional descriptor space such as 2D fingerprints. The approach used in BKD is the Parzen windows approach.

KDE can be chosen to estimate the likelihood of active or inactive molecules separately. It is convenient that the kernel function is itself a density probability mass function but not necessary when only the likelihood ratio is required. Hence, for active molecule selection purposes, a scoring function, is calculated from the likelihood ratio of the estimated distributions of active and inactive molecules which are estimated by Parzen windows. The higher the score, the more likely a molecule is to be active.

[83] compared BKD with merged similarity search and feedforward NNs. BKD can perform robustly with varying quantities of training data and also in the presence of noisy and sparse data. BKD was compared to other ranking methods for VS: similarity methods, trend vectors, substructural analysis and bioactivity profiles by [17]. These methods were tested on the NCI AIDS database and the Syngenta

corporate database[9]. The study found that BKD yielded consistently superior rankings and would appear to have considerable potential for chemical screening applications.

[84] investigated three different ways of carrying out VS when multiple bioactive reference structures are available.

1) Merging the individual fingerprints into a single combined fingerprint.
2) Applying data fusion to the similarity rankings resulting from individual similarity search.
3) Approximations to substructural analysis. The experiment used the MDDR database, [84] suggested that fused similarity scores are the most effective general approach with the best individual results coming from the BKD technique.

The study of the BKD method was extended with a comparison of a range of different types of 2D fingerprints [85]. The experiment showed that the ECFP_4 fingerprint [72] should be considered as a first choice as it can achieve a better overall performance.

Author's previous work [86] compare BKD, kernel Fisher discriminant analysis, kernel logistic regression, and their variants together. The results show that BKD in conjunction of Jaccard/Tanimoto performs the best in both homogeneous and heterogeneous classes. Moreover, BKD does not provide sparse solutions. This is important as the speed of recall is an important issue in VS tasks. The results show that sparse classifiers are competitive to the modification of BKD in most homogeneous classes, on the other hand, they are generally worse in most heterogeneous classes. This is because a sparse solution might lose some significant information in Gram matrix in heterogeneous classes.

Recently, [87] introduced the continuous kernel discrimination (CKD) which is based on the idea of original Parzen windows [81]. Gaussian radial basis function (GRBF) was used as the kernel function which is suitable for binary, integer, and real-valued representation of molecule structure. Their method was applied to 11 activity classes and 8 most diverse activity classes from the MDDR database. The database is represented as three non-binary descriptors and one binary descriptor The first type of non-binary descriptor was generated with the SciTegic Pipeline Pilot software [88] and contains integer and real values. The second descriptor, a 997-element integer vector, was generated with SYBYL holograms [89], and the third by Molconn-Z descriptors [90], a set of topological indices of molecular structure that have been used extensively for QSAR. The binary descriptor used is the ECFP_4 fingerprint. Their results show that the ECFP_4 fingerprint is the best fingerprint followed by holograms, Pipeline Pilot, and Molconn - Z representations. They compared CKD with BKD, the average active molecules retrieved by BKD is 79.7% while CKD achieves 78.1%. Thus, CKD is competitive to BKD as differences are not great. More recent work on CKD can be found from [91].

[9] The Syngenta corporate database contains 132,784 molecules which have been tested in various *in vivo* whole organism screens.

4.3.7. Graph Kernel Machines

As mentioned previously, the alternative representations of molecules are 2D or 3D graphs representations. Usually chemical structures are represented as *molecular graph* [7]. They are considered to be graphs in mathematical terms. Graphs consist of *nodes* and *edges*. In chemical analogy, nodes correspond to atoms, and edges are bonds. Graph theory is a well researched area of mathematics with applications in a wide range of disciplines.

Recently, [92], and [93] introduced positive definite kernels between labelled graphs which are based on the detection of common fragments between different graphs. However, these graph kernels are faced with two main problems: (i) the problem of computational complexity which is in proportion to the product of the size of two compared graphs, and (ii) the use of all fragments to characterize each graph might not be optimal because many fragments are irrelevant as they are represented by *tottering paths*[10] on the graph.

To address the above issues and enhance predictive accuracy, [94] proposed two extensions of the original graph kernel. The first extension is to re-label each vertex automatically in order to insert information about the environment of each vertex in its label. The second extension is to modify the random walk model which is proposed by [93]. These two modifications are used in conjunction with the SVM. Their experiments were conducted on two mutagenicity datasets [95, 96]. Their results show an improvement on area under receiver operating characteristic curve of the predictions and computation times. They compared their proposed algorithm with linear regression, decision tree, NNs, and inductive logic programming on a mutagenicity dataset [95]. They used leave-one-out cross-validation to perform their algorithm. It can achieve an accuracy of 88.1% which outperforms other compared methods.

[97] reviewed the literature on graph kernels and also introduce three new kernel functions to use with graphs namely: Tanimoto, MinMax, and Hybrid. These kernel functions are applied to three classification problems. Their results are at least comparable with or often better than previous results. The proposed method can achieve accuracy of 91.5% on the mutagenicity dataset [95], 65-67% on the Predictive Toxicology Challenge dataset [98], and 72% on the NCI Cancer Cell Lines dataset[11].

4.4. Docking

Docking aims to predict the 3D structure formed when one or more molecules form an intermolecular complex (fitting a molecule into a protein). Figure 4 shows an example of

[10] The path immediately returns to a visited vertex after leaving it.

[11] The NCI Cancer Cell Lines database is available from NCI/NIH Development Therapeutics Programme at URL http://dtp.nci.nih.gov/docs/misc/common_files/cell_list.html. It contains screening results for the ability of roughly 70,000 compounds to kill or inhibit the growth of a panel of 60 human tumour cell lines.

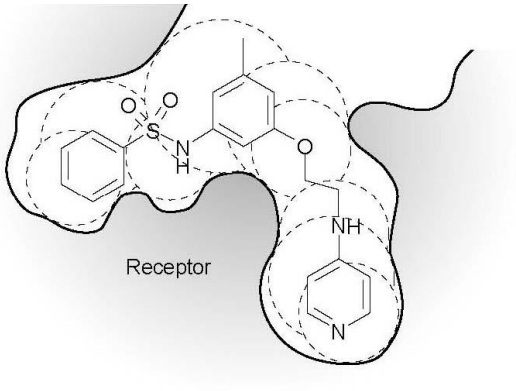

Figure 4. An exampleofdocking.

docking. The process consists of two components: (i) mechanism for exploring the coordinate space of the binding site (possible protein-ligand geometries – sometimes called *poses*) and (ii) scoring each possible ligand pose, which is then taken as the predicted binding mode for that compound.

Many algorithms have been proposed for protein docking. They differ in the handling of protein flexibility and scoring function. The first algorithm for docking was called DOCK [99]. It focuses on shape complementarity which is represented by a sphere-based description of the geometries of the receptor and the potential ligands. Recently, DOCK was extended to build up multiple conformations of each of the ligands from fragments inside the protein binding side [100]. Gold [101] and AutoDock [102] make use of genetic algorithms to perform a conformational search and dock the potential ligands. More reviews of docking algorithms can be found in [103, 7].

5. CONCLUSION

VS is a very important integral part of the drug discovery process and is effective in improving its efficiency. It is implemented as an iterative scheme. Many methods have been introduced and reintroduced. The methods used in VS tasks have been reviewed in this paper. The review emphasises machine learning. As it has become crucial as computers are expected to solve increasingly complex problems. Moreover it become more integrated into our daily lives. Current research efforts in VS algorithms focus on improving of accuracy rate and computational time. Hence, optimization is an important part of VS algorithms. There are many machine learning techniques introduced these day. Hence, it is hard to point which algorithm outperforms the others and is the best in VS. [59] compared a set of machine learning techniques and pointed out their advantages and disadvantages. There are also some works which compare a set of algorithms on the same benchmark – MDDR Database i.e. [86, 87, 104]. Machine learning algorithms are now playing a role not only in VS tasks but also in other real-world applications. The author believes that they will become a main role in VS tasks in coming years because machine learning techniques emphasize on obtaining accurate predictions and there is also a growing literature on machine learning in VS in the past years.

REFERENCES

[1] K. M. De Cock, "The eradication of smallpox: Edward Jenner and the first and only eradication of a human infectious disease," *Nature Medicine*, vol. 7, pp. 15–16, 2001.

[2] P. Scott, "Edward Jenner and the discovery of vaccination," 1999, accessed on 26 March 2012. [Online]. Available: http://www.sc.edu/library/spcoll/nathist/jenner.html

[3] M. Bellis, "The history of penicillin," 2007, accessed on on 26 March 2012. [Online]. Available: http://inventors.about.com/od/pstartinventions/a/Penicillin.htm

[4] J. Drews, *In Quest of Tomorrows Medicines. Springer*, 1999.

[5] A. De Costanzo, "How do drugs work?" 2007, accessed on 26 March 2012. [Online]. Available: http://drugs.about.com/od/azdrugsbycondition/a/how_drugs_work.htm

[6] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: New estimates of drug development costs," *Journal of Health Economics*, vol. 22, pp. 151–185, 2003.

[7] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*. Dordrecht: Kluwer Academic Publishers, 2003.

[8] V. L. Nienaber, P. L. Richardson, V. Klighofer, J. J. Bouska, V. L. Giranda, and J. Greer, "Discovering novel ligands for macromolecules using Xray crystallographic screening," *Nature Biotechnology*, vol. 18, pp. 1105–1108, 2000.

[9] J. Rogers, "Mouse clues to human genetics," 2002, accessed on 26 March 2012. [Online]. Available: http://news.bbc.co.uk/1/hi/sci/tech/2536501.stm

[10] American Association for Laboratory Animal Science, "The drug discovery process," 2004, world Wide Web, http://www.aalas.org/doc/sect-1_4.doc. [Online]. Available: http://www.aalas.org

[11] Syagen Technology Inc., "Bioanalytical," 2007, accessed on 26 March 2012. [Online]. Available: http://www.syagen.com/bioanalytical.asp

[12] F. Brown, "Chemoinformatics: What is it and how does it impact drug discovery," *Annual Reports in Medicinal Chemistry*, vol. 33, pp. 375–384, 1998.

[13] M. Hann and R. Green, "Chemoinformatics – a new name for an old problem?" *Current Opinion in Chemical Biology,* vol. 3, no. 4, pp. 379–383, 1999.

[14 T. Oprea, Chemoinformatics : *Concepts, methods, and tools for drug discovery*. New York: Wiley-VCH, 2005.

[15] H. Böhm and G. Schneider, *Virtual Screening for Bioactive Molecules*. New York, USA: John Wiley & Sons, Inc., 2000.

[16] W. Walters, M. Stahl, and M. Murcko, "Virtual screening – an overview," *Drug Discovery Today*, vol. 3, no. 4, pp. 160–178, 1998.

[17] D. Wilton, P. Willett, K. Lawson, and G. Mullier, "Comparison of ranking methods for virtual screening in lead-discovery programs," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 469–474, 2003.

[18] R. Carhart, D. Smith, and R. Venkataraghavan, "Atom pairs as molecular features in structureactivity studies: definition and applications," *Journal of Chemical Information*

and Computer Sciences, vol. 25, no. 2, pp. 64–73, 1985.

[19] P. Willett, V. Winterman, and D. Bawden, "Implementation of nearest-neighbor searching in an online chemical structure search system," *Journal of Chemical Information and Computer Sciences*, vol. 26, no. 1, pp. 36–41, 1986.

[20] J. Raymond, E. Gardiner, and P. Willett, "RASCAL: Calculation of graph similarity using maximum common edge subgraphs," *The Computer Journal*, vol. 45, no. 6, pp. 631–644, 2002.

[21] D. Vidal, M. Thormann, and M. Pons, "LINGO – an efficient holographic text based method to calculate biophysical properties and intermolecular similarities," *Journal of Chemical Information and Modeling*, vol. 45, no. 2, pp. 386–393, 2005.

[22] M. A. Johnson and G. M. Maggiora, Eds., *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons, 1990.

[23] Y. C. Martin, J. L. Kofron, and L. Traphagen, "Do structurally similar molecules have similar biological activity?" *Journal of Medicinal Chemistry*, vol. 45, no. 19, pp. 4350–4358, 2002.

[24] D. C. I. S. Inc., "Daylight fingerprints," 2006,

world Wide Web, http://www.daylight.com/. [Online]. Available: http://www.daylight.com/

[25] J. C. Gower, "Measures of similarity, dissimilarity and distance," *Encyclopedia of statistical sciences*, vol. 5, pp. 397–405, 1985.

[26] D. Ellis, J. Furner-Hines, and P. Willett, "Measuring the degree of similarity between objects in text retrieval systems," *Perspectives in Information Management*, vol. 3, no. 2, pp. 128–149, 1993.

[27] P. Willett and V. Winterman, "A comparison of some measures of intermolecular structural similarity," *Quantitative Structure-Activity Relationships*, vol. 5, pp. 18–25, 1986.

[28] J. Raymond and P. Willett, "Maximum common subgraph isomorphism algorithms for the matching of chemical structures," Journal *of Computer-Aided Molecular Design*, vol. 16, no. 7, pp. 521–533, 2002.

[29] T. Hagadone, "Molecular substructure similarity searching: Efficient retrieval in two-dimensional structure databases," *Journal of Chemical Information and Computer Sciences*, vol. 32, no. 5, pp. 515–521, 1992.

[30] J. Raymond and P. Willett, "Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases," *Journal of Computer-Aided Molecular Design*, vol. 16, no. 1, pp. 59–71, 2002.

[31] D. Vidal, M. Thormann, and M. Pons, "A novel search engine for virtual screening of very large databases," *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 836–843, 2006.

[32] C. M. Ginn, P. Willett, and J. Bradshaw, "Combination of molecular similarity measures using data fusion," *Perspectives in Drug Discovery and Design*, vol. 20, pp. 1–16, 2000.

[33] N. Belkin, P. Kantor, E. Fox, and J. Shaw, "Combining the evidence of multiple query representations for information retrieval," *Information Processing and Management*, vol. 31, no. 3, pp. 431–448, 1995.

[34] N. Salim, J. Holliday, and P. Willett, "Combination of fingerprint-based similarity coefficients using data fusion," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 435–442, 2003.

[35] J. Holliday, C. Hu, and P. Willett, "Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings," *Combinatorial Chemistry and High-Throughput Screening*, vol. 5, no. 2, pp. 155–166, 2002.

[36] MDL Information Systems Inc., "The MDL drug data report database," 2006, world Wide Web, http://www.mdli.com. [Online]. Available: http://www.mdli.com

[37] R. Wang and S. Wang, "How does consensus scoring work for virtual library screening? An idealized computer experiment," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 5, pp. 1422–1426, 2001.

[38] P. S. Charifson, J. J. Corkery, M. A. Murcko, and P. W. Walters, "Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins," *Journal of Medicinal Chemistry*, vol. 42, pp. 5100–5109, 1999.

[39] A. Klon, M. Glick, and J. Davies, "Combination of a naive bayes classifier with consensus scoring improves enrichment of high-throughput docking results," *Journal of Medicinal Chemistry*, vol. 47, no. 18, pp. 4356–4359, 2004.

[40] O. Güner, Ed., *Pharmacophore Perception, Development and Use in Drug design*. California: International University Line, 2000.

[41] C. G. Wermuth, C. R. Ganellin, P. Lindberg, and L. A. Mitscher, "Glossary of terms used in medicinal chemistry (iupac recommendations 1998)," *Pure & Applied Chemistry*, vol. 70, no. 5, pp. 1129–1143, 1998.

[42] W. A. Warr and P. Willett, "The principles and practice of 3D database searching," *in Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*, P. Willett and Y. Martin, Eds. Washington D.C.: American Chemical Society, 1997, pp. 77–95.

[43] Y. C. Martin, "Distance comparisons: A new strategy for examining three-dimensional structureactivity relationships," *Classical and Three-Dimensional QSAR in Agrochemistry*, vol. 606, pp. 318–329, 1995.

[44] G. Jones, P. Willett, and R. C. Glen, "A genetic algorithm for flexible molecular overlay and pharmacophore elucidation," *Journal of Computer-Aided Molecular Design*, vol. 9, no. 6, pp. 532–549, 1995.

[45] T. Langer and R. D. Hoffmann, Eds., *Pharmacophores and Phamacophore Searched*. Weinheim: Wiley-VCH, 2006.

[46] R. D. Cramer, G. Redl, and C. E. Berkoff, "Substructural analysis. A novel approach to the problem of drug design," *Journal of Medicinal Chemistry*, vol. 17, pp. 533–535, 1974.

[47] A. Ormerod, P. Willett, and D. Bawden, "Comparison of fragments weighting schemes for substructural analysis," *Quantitative Structure-Activity Relationships*, vol. 8, pp. 115–129, 1989.

[48] D. A. Cosgrove and P. Willett, "SLASH: A program for analysing the functional groups in molecules," *Journal of Molecular Graphics and Modelling*, vol. 16, no. 1, pp. 19–32, 1998.

[49] A. M. Capelli, A. Feriani, G. Tedesco, and A. Pozzan, “Generation of a focused set of GSK compounds biased toward ligand-gated ion-channel ligands,” *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 659–664, 2006.

[50] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, “New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching,” *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 462–470, 2006.

[51] Y. Martin, J. Holland, C. Jarboe, and N. Plotnikoff, “Discriminant analysis of the relationship between physical properties and the inhibition of monoamine oxidase by aminotetralins and aminoindans,” *Journal of Medicinal Chemistry*, vol. 17, pp. 409–413, 1974.

[52] J. Prieto, A. Talevi, and L. E. Bruno-Blanch, “Application of linear discriminant analysis in the virtual screening of antichagasic drugs through trypanothione reductase inhibition,” *Molecular Diversity*, vol. 10, no. 3, pp. 361–375, 2006.

[53] H. González-Díaz, O. Gia, E. Uriarte, I. Hernández, R. Ramos, M. Chaviano, S. Seijo, J. Castillo, L. Morales, L. Santana, D. Akpaloo, E. Molina, M. Cruz, L. Torres, and M. Cabrera, “Markovian chemicals “in silico”design MARCH-INSIDE, a promising approach for computer-aided molecular design I: Discovery of anticancer compounds,”*Journal of Molecular Modeling*, vol. 9, pp. 395–407, 2003.

[54] N. Mahmoudi, J.-V. Julian-Ortiz, L. Ciceron, J. Galvez, D. Mazier, M. Danis, F. Derouin, and R. Garcia-Domenech, “Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening,” *Journal of Antimicrobial Chemotherapy*, vol. 57, pp. 489–497, 2006.

[55] J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, 2nd ed. New York, USA: John Wiley & Sons, Inc., 1999.

[56] S. Fujishima and Y. Takahashi, “Classification of dopamine antagonists using TFS-based artificial neural network,” *Journal of Chemical Information and Modeling*, vol. 44, no. 3, pp. 1006–1009, 2004.

[57] R. Mueller, E. S. Dawson, J. Meiler, A. L. Rodriguez, B. A. Chauder, B. S. Bates, A. S. Felts, J. P. Lamb, U. N. Menon, S. B. Jadhav, A. S. Kane, C. K. Jones, K. J. Gregory, C. M. Niswender, P. J. Conn, C. M. Olsen, D. G. Winder, K. A. Emmitte, and C. W. Lindsley, “Discovery of 2-(2-Benzoxazoyl amino)-4-Aryl-5-Cyanopyrimidine as negative allosteric modulators (NAMs) of metabotropic glutamate receptor 5 mGlu5): From an artificial neural network virtual screen to an in vivo tool compound,” ChemMed- Chem, vol. 7, no. 3, pp. 406–414, 2012.

[58] S. Bandholtz, J. Wichard, R. Kühne, and C. Grötzinger, “Molecular evolution of a peptide gpcr ligand driven by artificial neural networks,” PLoS ONE, vol. 7, no. 5, p. e36948, 05 2012.

[59] I. Muegge and S. Oloff, “Advances in virtual screening,” *Drug Discovery Today: Technologies*, vol. 3, no. 4, pp. 405–411, 2006.

[60] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, pp. 81–106, 1986.

[61] ——, *C4.5 Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers Inc., 1993.

[62] F. Cheng, Y. Ikenaga, Y. Zhou, Y. Yu, W. Li, J. Shen, Z. Du, L. Chen, C. Xu, G. Liu, P. W. Lee, and Y. Tang, "In silico assessment of chemical biodegradability," *Journal of Chemical Information and Modeling*, vol. 52, no. 3, pp. 655–669, 2012.

[63] D. Plewczynski, S. Spieser, and U. Koch, "Assessing different classification methods for virtual screening," Journal of Chemical Information and Modeling, vol. 46, no. 3, pp. 1098–1106, 2006.

[64] L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider, and G. Schneider, "Extraction and visualization of potential pharmacophore points using support vector machines: Application to ligand-based virtual screening for COX-2 inhibitors," *Journal of Medicinal Chemistry*, vol. 48, no. 22, pp. 6997–7004, 2005.

[65] K. Müller, G. Rätsch, S. Sonnenburg, S. Mika,M. Grimm, and N. Heinrich, Classifying 'druglikeness' with kernel-based learning methods," *Journal of Chemical Information and Modeling*, vol. 45, no. 2, pp. 249–253, 2005.

[66] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1882–1889, 2003.

[67] C. Leslie, E. Eskin, and W. Noble, "The spectrum kernel: A string kernel for SVM protein classification," in *Proceeding of the Pacific Symposium on Biocomputing*, 2–7 January 2002, pp. 474–485.

[68] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, USA: Springer-Verlag New York, Inc., 1995.

[69] T. Jaakkola, M. Diekhans, and D. Haussler, "A discriminative framework for detecting remote protein homologies," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 95–114, 2000.

[70] R. N. Jorissen and M. K. Gilson, "Virtual screening of molecular databases using a support vector machine," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 549–561, 2005.

[71] C.-C. Chang and C.-J. Lin, LIBSVM: *A library for support vector machines*, 2001, world Wide Web, http://www.csie.ntu.edu.tw/_cjlin/libsvm/.

[72] S. Inc., "ECFP_4 fingerprints," 2006, world Wide Web, http://www.scitegic.com/. [Online]. Available: http://www.scitegic.com/

[73] Q. Li, A. Bender, J. Pei, and L. Lai, "A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification," *Journal of Chemical Information and Modeling*, vol. 47, no. 5, pp. 1776–1786, 2007.

[74] S. Bhavani, A. Nagargadde, A. Thawani, V. Sridhar, and N. Chandra, "Substructure-based support vector machine classifiers for prediction of adverse effects in diverse classes of drugs," *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2478–2486, 2006.

[75] L. Terfloth, B. Bienfait, and J. Gasteiger, "Ligandbased models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates," *Journal of Chemical*

Information and Modeling, vol. 47, no. 4, pp. 1688–1701, 2007.

[76] L.-J. Tang, Y.-P. Zhou, J.-H. Jiang, H.-Y. Zou, H.-L. Wu, G.-L. Shen, and R.-Q. Yu, "Radial basis function network-based transform for a nonlinear support vector machine as optimized by a particle swarm optimization algorithm with application to QSAR studies," *Journal of Chemical Information and Modeling*, vol. 47, no. 4, pp. 1438–1445, 2007.

[77] F. Rathke, K. Hansen, U. Brefeld, and K.-R. Muller, "StructRank: A new approach for ligandbased virtual screening," *Journal of Chemical Information and Modeling*, vol. 51, no. 1, pp. 83–92, 2011.

[78] L. Wang, C. Ma, P. Wipf, and X.-Q. Xie, "Linear and nonlinear support vector machine for the classification of human 5-HT1A ligand functionality," *Molecular Informatics*, vol. 31, no. 1, pp. 85–95, 2012.

[79] H.-L. Wan, Z.-R. Wang, L.-L. Li, C. Cheng, P. Ji, J.-J. Liu, H. Zhang, J. Zou, and S.-Y. Yang, "Discovery of novel bruton's tyrosine kinase Btk inhibitors using a hybrid protocol of virtual screening approaches based on svm model, pharmacophore, and molecular docking," *Chemical Biology & Drug Design*, pp. no–no, 2012.

[80] J. Aitchison and C. G. Aitken, "Multivariatebinary discrimination by the kernel method," *Biometrika*, vol. 63, no. 3, pp. 413–420, 1976.

[81] E. Parzen, "On the estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

[82] G. Harper, "The selection of compounds for screening in pharmaceutical research," Ph.D. dissertation, The University of Oxford, 1999.

[83] G. Harper, J. Bradshaw, J. C. Gittins, D. V. Green, and A. R. Leach, "Prediction of biological activity for high-throughput screening using binary kernel discrimination," *Journal of Chemical Information and Computer Sciences*, vol. 41, pp. 1295–1300, 2001.

[84] P. Hert, J.and Willett, D. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, "Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures," *Journal of Chemical Information and Modeling*, vol. 44, no. 3, pp. 1177–1185, 2004.

[85] J. Hert, P. Willett, D. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, "Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures," *Organic & Biomolecular Chemistry*, vol. 2, no. 22, pp. 3256–3266, 2004.

[86] K. Pasupa, "Data mining and decision support in pharmaceutical databases," Ph.D. dissertation, Department of Automatic Control & Systems Engineering, University of Sheffield, 2007.

[87] B. Chen, R. Harrison, G. Papadatos, P. Willett, D. Wood, X. Lewell, P. Greenidge, and N. Stiefl, "Evaluation of machine-learning methods for ligand-based virtual screening." *Journal of Computer-Aided Molecular Design*, vol. 21, pp. 53–62, 2007.

[88] S. Inc., "SciTegic Pipeline Pilot software," 2006, world Wide Web, http://www.scitegic.com/. [Online]. Available: http://www.scitegic.com/

[89] T. Inc., "SYBYL holograms," 2006, world Wide Web, http://www.tripos.com/. [Online]. Available: http://www.tripos.com/

[90] eduSoft LC, "Molconn-Z descriptors," 2006, world Wide Web, http://www.edusoftlc.com/.[Online]. Available: http://www.edusoftlc.com/

[91] J. He, G. Yang, H. Rao, Z. Li, X. Ding, and Y. Chen, "Prediction of human major histocompatibility complex class II binding peptides by continuous kernel discrimination method," *Artificial Intelligence in Medicine*, vol. 55, no. 2, pp.107–115, 2012.

[92] H. Kashima, K. Tsuda, and A. Inokuchi, "Marginalized kernels between labeled graphs," *in Proceedings of the Twentieth International Conference on Machine Learning* (ICML'2003), 2003, pp. 321–328.

[93] T. Gärtner, P. Flach, and S. Wrobel, "On graph kernels: Hardness results and efficient alternatives," *in Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*. Springer-Verlag, August 2003, pp. 129–143.

[94] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J. Vert, "Graph kernels for molecular structureactivity relationship analysis with support vector machines," *Journal of Chemical Information and Modeling*, vol. 45, no. 4, pp. 939–951, 2005.

[95] A. K. Debnath, R. L. Lopezđe Compadre , G. Debnath, A. J. Shusterman, and C. Hansch, "Structureactivity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity," *Journal of Medicinal Chemistry*, vol. 34, no. 2, pp. 786–797, 1991.

[96] C. Helma, T. Cramer, S. Kramer, and L. DeRaedt, "Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds," *Journal of Chemical Information and Modeling*, vol. 44, no. 4, pp. 1402–1411, 2004.

[97] L. Ralaivola, S. Swamidass, H. Saigo, and P. Baldi, "Graph kernels for chemical informatics," *Neural Networks*, vol. 18, no. 8, pp. 1093–1110, 2005.

[98] C. Helma, R. D. King, S. Kramer, and A. Srinivasan, "The predictive toxicology challenge 2000–2001," *Bioinformatics*, vol. 17, no. 1, pp. 107–108, 2001.

[99] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, "A geometric approach to macromolecule-ligand interactions," *Journal of Molecular Biology*, vol. 161, no. 2, pp. 269–288, 1982.

[100] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases," *Journal of Computer-Aided Molecular Design*, vol. 15, no. 5, pp. 411–428(18), 2001.

[101] G. Jones, P. Willett, and R. C. Glen, "Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation," *Journal of Molecular Biology*, vol. 245, no. 1, pp. 43–53, 1995.

[102] G. M. Morris, D. S. Goodsell, R. S. Halliday,R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson,"Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function," *Journal of Computational*

Chemistry, vol. 19, no. 14, pp. 1639–1662, 1998.

[103] P. D. Lyne, "Structure-based virtual screening: An overview," *Drug Discovery Today*, vol. 7, no. 20, pp. 1047–1055, 2002.

[104] B. Chen, R. F. Harrison, K. Pasupa, P. Willett, D. J. Wilton, D. J. Wood, and X. Q. Lewell, "Virtual screening using binary kernel discrimination: Effect of noisy training data and the optimization of performance." *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 478–486, 2006.