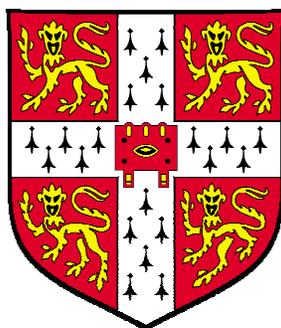


Andreas Bender  
Darwin College

RESEARCH PROJECT

**NEW APPROACHES TO MOLECULAR SIMILARITY**

CPGS DISSERTATION



University of Cambridge  
Unilever Centre for Molecular Science Informatics  
Department of Chemistry

Supervisor  
Robert C Glen

2003

## PREFACE

This work was carried out by Andreas Bender from January to November 2003 under supervision of Prof Robert C Glen at the Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge. The work contained in this dissertation, or any part thereof, has not been submitted for any other degree. This thesis contains 11 352 words in total.

### Acknowledgements

I would like to express my gratitude to my supervisor, Bobby Glen, for his help and valuable advice during this project. Hamse Y. Mussa and James Smith gave me the possibility for many helpful discussions. The Gates Cambridge Trust and Unilever are thanked for funding.

### Reference

The experimental part of this dissertation has been accepted for publication by the *Journal of Chemical Information and Computer Sciences*.

### Copyright

Copyright 2003 Andreas Bender

## **SUMMARY**

A novel technique for similarity searching is introduced. Molecules are represented by atom environments, which are fed into an information-gain based feature selection. A Naïve Bayesian Classifier is then employed for compound classification. The new method is tested by its ability to retrieve five sets of active molecules seeded in the MDL Drug Data Report (MDDR). In comparison experiments, the algorithm outperforms all current retrieval methods which use two- and three-dimensional descriptors and offers insight into the significance of structural components for binding.

## CONTENTS

PREFACE .....	2
SUMMARY .....	3
CONTENTS.....	4
1. INTRODUCTION .....	5
2. LITERATURE REVIEW.....	7
a) Representation of Molecules.....	8
b) Comparison of Molecules.....	13
3. EXPERIMENTAL DETAILS .....	16
a) Descriptor Generation / Molecular Representation.....	16
b) Feature Selection .....	17
c) Classification .....	18
d) Compilation of Dataset and Pre-processing.....	19
e) Calculations.....	20
4. RESULTS.....	22
5. DISCUSSION .....	36
6. CONCLUSIONS .....	40
7. FUTURE WORK.....	41
REFERENCES.....	42

## 1. INTRODUCTION

The question of how to describe similarity of molecules has become increasingly important over the last two decades and is likely to become even more important in the future. There are a number of reasons for this tendency.

According to the 2003 report by the Tufts Center for Drug Development [DiMasi 2003], costs of a single new chemical compound until the point of submission to approval has risen to US\$ 802 million. This is due to high failure rates in later stages of drug development. Probably the “easiest cherries have already been picked” - drugs for easily tractable targets have already been found. Furthermore it is well known that *in vitro* and *in vivo* screenings are very expensive, compared to so-called *in silico* approaches.

Another major reason for the surge in similarity searching is the negative public opinion with respect to animal testing, so much that this results in its ban in home and personal care products in the European Union starting from 2009 [Europarl 2003].

However, computers have become much more powerful and cheaper over the last years, thereby allowing *in silico* screening using larger databases and more sophisticated algorithms. Using appropriate similarity measures, it might become possible to predict properties like absorption, distribution, metabolism, excretion or toxicity (ADME/Tox) at an earlier stage of the research pipeline, reducing expenditure per successful compound [van de Waterbeemd 2003]. Only the most promising compounds will then be synthesized and screened, potentially yielding a higher fraction of active structures in the selected subset and higher survival rates.

In order to avoid animal testing, cosmetics and other consumer goods companies will focus on their in-house databases of chemical compounds that have already been tested for safety. Out of these compounds, some of them might already possess the desired properties, which could be detected by similarity searching.

The rest of the thesis is structured as follows. In section 2, a review of the literature on molecular representation and molecular similarity searching is given. Section 3 presents experimental details. Results are given in section 4, which are discussed fully

in section 5. We draw conclusions in section 6 and outline possible further research directions in section 7.

## 2. LITERATURE REVIEW

Similarity searching is based on the “Similar Property Principle” [Johnson 1991] that states that structurally similar molecules - structures with a “similar” spatial arrangement of “similar” functional groups - tend to have similar properties, physical as well as biological ones. All current drug design efforts are based on this paradigm.

Similarity is a concept that is present in everyday life, e.g. in visual perception, and has thus been subject to intensive psychological research [Tversky 1977]. Many of the ideas behind similarity measures currently employed in comparison of molecules are rooted in psychology. An illustration of asymmetrical perception of similarity was given by Tversky [Tversky 1977]. He was asking whether North Korea was more similar to China, or that China was more similar to North Korea. A consistent answer (the former option of the two) was given by analysis of his test subjects; consistent with the ubiquitous finding that one representative of a class is also usually found to be more similar to the class than the class being similar to the member. This illustrates the origin of asymmetrical similarity measures. Rouvray gives a comprehensive overview of similarity applications in the natural sciences [Rouvray 1992].

The definition of similarity with respect to molecules is more stringent than that in other fields. Basically it consists of mapping “chemical space” (a representation of a molecule in structural or some property space) to one-dimensional space with entities of real numbers. Ideally similarity measures for molecules behave proportionally to all physical and biological properties of molecules in this representation. In other words, it groups together all molecules with very similar physical and biological properties in a confined area of chemical property space. In practice, we are far away from reaching this goal. As we will see in the following paragraphs, molecular representations have to this day only been applied to specific problems of molecular similarity.

Similarity searches complement earlier substructure searches [Hagadone 1992] which only consider presence or absence of specific features but did not evaluate global properties and overall shape. Compared to substructure searches, similarity searches

are both more general and more comprehensive. They are more general by employing abstract representations of molecules or molecular properties and by being capable of using fuzzy matching techniques. Furthermore they are more comprehensive as they (usually) comprise features derived from the whole molecule under consideration.

Molecular similarity calculations are done in three steps: representation of the molecules in descriptor space, feature selection, and comparison. The literature review in the following paragraphs will focus mainly on representation and comparison of molecules.

### **a) Representation of Molecules**

A variety of methods to represent molecules in chemical space are known. Here we divide them into one-dimensional descriptors, topological indices, fragment-based descriptors, field-based descriptors, subshape descriptors, surface-derived descriptors, affinity fingerprints, spectra-derived descriptors, and back-projectable descriptors.

The first group of descriptors give one-dimensional property descriptors or one-dimensional linear representations of the whole molecule. One-dimensional property descriptions assign only one number to the molecule. This number is usually derived from physicochemical properties. This provides the basis for variable selection structure-activity regression techniques. Since no geometrical information is contained in the descriptor, they are most commonly employed for the prediction of physical properties as opposed to receptor binding. Good examples using this descriptor are clustering of compound databases [Downs 1994] and database comparisons (distinguishing between drugs/non-drugs [Jain 1998, Lipinski 1997, Lipinski 2000]).

One-dimensional linear representations attempt to represent the molecule as a linear tree where nodes represent atoms (or groups of atoms). This is similar to the representation of proteins in one-dimensional sequences of amino acids. To compare molecules, algorithms similar to protein sequence alignments can be applied to

compare two molecules [Dixon 2001]. An overview of methods to derive linear molecular descriptors is given in [Baumann 1999].

Topological indices and other graph-based descriptors constitute the second group of descriptors. Topological indices are integer or real-valued numbers that are derived from the connectivity matrix and sometimes they contain additional property information of the molecule. They are generally divided into three generations of indices. The first generation, such as the Wiener index, are derived from integer graph properties and are themselves integers. Second generation indices, such as the molecular connectivity indices, are real numbers derived from integer graph properties whereas indices of the third generation are real valued numbers derived from real valued graph properties. Several hundred alternative topological descriptors have been published to this day [Wilkins 1980, Randic 1979; Balaban 1982]. One important aspect of topological indices is that they are derived solely from the connectivity matrix of a molecule and thus do not consider both conformations and three-dimensional structure. For a recent review on topological indices, see [Balaban 1995] and [Estrada 2001].

The next group of descriptors are fragment or substructure based descriptors. Maximum common substructure (MCS) searches are among the earliest substructure searching algorithms used [Cone 1977]. These searches tend to be time-consuming due to the NP-complete nature of the problem which in the worst scenario becomes exhaustive. Recent advances can be found in [Barnard 1993]. Substructural analysis is often dubbed Free-Wilson-Analysis as Free and Wilson published one of the early works in this area [Free 1964, Cramer 1974]. It has been an active area ever since as more recent publications show [Gillet 1998]. Kier and Hall [Hall 1995] extended topological descriptors to include electronic and valence state information in their “electrotopological” descriptors, an approach that has later been extended to “E-state fields” [Kellogg 1996]. Rarey and co-workers [Rarey 1998] represent molecules as one-dimensional, potentially branched, sequences which they called “Feature Trees”. Other examples for fragment-based descriptors are [Takahashi 1992, Barnard 1993] using reduced graphs, [ Faulon 1994, Faulon 2003a, Faulon 2003b, Visco Jr. 2002], using “molecular tree” fingerprints and [Xing 2002, Xing 2003, Bender 2004] using related “Atom Environments”. “Mini fingerprints” also contain bits which denote the

presence or absence of fragments [Xue 2002, Xue 2001, Xue 1999]. A review of fragment-based measures of molecular similarity is given in [Bath 1994] which finds that a description using four-atom fragments is most effective.

The group of field-based descriptors differs from the previous group in that they use three-dimensional information of a molecule for their derivation. Because of the number of data points (“grid points”) that are necessary for a sensible resolution, they are computationally more demanding than two-dimensional methods. Field-based descriptors generally require alignment of the molecules to be compared that is only trivial in case of analogue compounds. Many different methods have been developed in this area with the broad separation being between quantum-mechanical methods and non-quantum mechanical methods. Quantum Similarity has been introduced in the early 80’s [Carbó 1980] and since then it has been subjected to intensive research. Hodgkin [Hodgkin 1987] later introduced a related index that took into account not only electron distribution (such as the Carbó index) but also electron density. Walker [Walker 1991] and Good [Good 1992, Good 1993] replaced the grid approach with a Gaussian approximation. This led to significant increase in performance. Furthermore it solved problems with local minima while performing molecular alignments. The Gaussian representation has later been generalized to describe molecular shape [Grant 1995]. For a review on quantum similarity, see [Carbo-Dorca 1998], for a basic introduction to the subject see [Carbo 1992]. On the other hand, non-quantum mechanical grid based descriptors have been introduced in the late 1980’s with the Comparative Molecular Field Method, CoMFA [Cramer 1988]. This method was also the basis of Klebe’s Comparative Molecular Similarity Analysis (CoMSIA) approach [Klebe 1994, Klebe 1998].

The (sub-)shape based descriptors group describe the shape of a molecule not in one fragment, but instead use several small features to describe the molecules and find related structures by “circumstantial evidence”. These methods are free from alignment problems and are usually realized with a bit string representation of features that suits computer treatment. They are often referred to as multiple-point-pharmacophores: two-point pharmacophores (2PP, [Sheirdan 1989, Good 1995b, Sheridan 1996]), which are known as atom pairs and represent all possible pairs of atoms in the molecule, three-point pharmacophores (3PP, [Gund 1977, Bemis 1992,

Nilakantan 1993, Pickett 1996, Mason 2001, Martin 1992]) which allow for a more detailed representation of interatomic distances, and four-point pharmacophores (4PP, [Mason 1999, Duca 2001]) which are able to distinguish between geometric isomers.

The surface-based group descriptors focuses on the commonly accepted assumption that ligand-receptor binding is mediated by the molecular surface, e.g. by the Van-der-Waals surface.

- Gaillard et al [Gaillard 1994] devised a method to describe molecular lipophilicity potential and validated it by predicting logP values.
- Stanton and Jurs [Stanton 1990] introduced the concept of “charged partial surface area structural descriptors” and derived descriptors describing surface charge from it.
- Jain’s Compass method [Jain 1994] is able to take several molecules and several conformations into account, but it needs a user-defined interacting pharmacophore guess. This approach has also been used for selecting library subsets in its extension called Icepick [Mount 1999], where several conformations of the molecules to be compared are calculated and the three-dimensional structures are docked into each other.
- Jain [Jain 2000] introduced the concept of “morphological similarity” which is defined as a Gaussian function of the differences in molecular surface distances of two molecules at weighted observation points on a uniform grid; compared to field-based methods, this method has the advantage that no alignment is necessary.
- A novel method for classifying similarity of molecules is performed by using hashkeys of the molecular surface, compared to a panel of reference compounds [Ghuloum 1999]. Applied to several data sets, the description is found to capture enough information for the prediction of ADME properties and target binding. Hash codes have already been applied in chemistry before [Ihlenfeldt 1994], but only for structure storage and not for structure-activity relationships.

The group of affinity-fingerprint based descriptors compare a ligand to a panel of reference receptors and scores each ligand by docking it into each receptor. The

resulting affinity vector can then be used to create a similarity index for the group of ligands among each other. This approach is computationally demanding, because every ligand molecule has to be docked against every reference receptor molecule. On the other hand, the “expertise of the receptor” is the crucial property for finding ligands *in vivo*, so that more meaningful results may be retrieved from this approach. *In vitro* fingerprints were first introduced by Kauvar [Kauvar 1995] and shortly afterwards followed by their *in silico* counterparts [Briem 1996, Lessel 2000]. The latter were for example employed in library design [Dixon 1998], for a recent review see [Briem 2000].

The group of spectra-derived descriptors uses a “natural” way to derive a one-dimensional representation of a molecule. X-ray and electron diffraction as well as infrared spectra have been used in this sense. The resulting spectra have to be converted into descriptor space, e.g. by calculating its zero crossings. The earliest work in this area was done by Soltzberg [Soltzberg 1976], who used molecular transforms to calculate the diffraction pattern from an X-ray derived three-dimensional structure. Electron diffraction was also used in the 3D-MoRSE (Molecule Representation of Structures based on Electron diffraction) approach [Schoonjans 1996]. The first descriptor calculated from the vibrational spectra of molecules is the EVA descriptor [Ginn 1997]. Here, fundamental frequencies of the vibrational spectrum are calculated and used for the comparison of molecules. A different approach [Schoonjans 2001] defines fuzzy peak areas to derive molecular features from an infrared spectrum, followed by principal component analysis. Although spectra are a “natural” way to convert a molecule into a one-dimensional representation, small changes often introduce major changes in the spectrum and the representation in descriptor space. These changes often make it difficult to use this approach as a similarity index.

The last and most recent group of molecular descriptors are the back-projectable descriptors. Those descriptors can be projected back on the molecules that were used to derive the descriptor in the first place and often hint at points where molecules can be optimised with respect to bioactivity. The first back-projectable descriptor was published by Pastor and co-workers [Pastor 2000] and was called GRIND (GRId INdependent Descriptors). First, a set of simplified molecular interaction fields

around the probe molecule is calculated. Commonly, a hydrophobic probe (DRY), an oxygen probe (O) and a nitrogen probe (N1) are used to distinguish between hydrophobic, hydrogen bond donor and hydrogen bond acceptor properties, respectively. In the second step, an alignment-independent descriptor based on autocorrelation is calculated. Another descriptor that falls into this area is the MaP (Mapping Property distributions of molecular surfaces) descriptor [Stiefl 2003]. This algorithm consists of three steps. Equally distributed surface points are computed first and then molecular properties are projected onto this surface. After that the distribution of surface points and properties is encoded into a translationally and rotationally invariant molecular descriptor which is based on radial distribution functions. An important feature of back-projectable descriptors is that they are easy to interpret.

### **b) Comparison of Molecules**

Comparison of molecules is usually performed using either similarity coefficients or machine learning approaches.

Several dozen similarity coefficients have been published. Similarity coefficients for the comparison of bit strings of molecules can be broadly divided into association, correlation and distance coefficients. Association coefficients try to capture fragments common to the two molecules to be compared and give a result in the range [0,1], where 1 represents identical molecules. The Tanimoto coefficient is an example of this class. Correlation coefficients give values in the same range and calculate the correlation between two vectors representing two molecules. The Pearson coefficient is a member of the class of correlation coefficients. Distance coefficients focus on differences between two molecules and are a measure of dissimilarity, giving results in the range [0, + inf]. One example is the Euclidean Distance.

Early work by Willett [Willett 1986] concluded that similarity calculation based on the Tanimoto coefficient on average performed best, when a total of 36 similarity coefficients were compared. A group of 22 different similarity coefficients has been

evaluated by Holliday [Holliday 2002], who found that some of the coefficients were exhibiting similar behaviour and that they could be grouped into several clusters. Hubalek [Hubalek 1982] lists 43 association coefficients, and found that 20 of those were synonymous to other coefficients. The remaining 23 coefficients were clustered into five groups.

On the other hand machine-learning approaches can be used to compare molecules.

Kernel Methods attempt to predict the output of a continuous output variable given continuous input variables. In drug-design, usually only the distinction between active and non-active entities is to be made. Then binary kernel methods are used, which can predict the output variable based on binary input vectors. One early publication on binary kernel methods was published by Aitchinson [Aitchison 1976]. He discusses the concept in general terms. More recently this concept has been revived by Harper [Harper 2001] and applied to a set of monoaminoxidase inhibitors.

Binary QSAR is related to binary kernel discrimination in that it also accepts binary input values (e.g., presence/absence of structural keys), but the kernel is exchanged for a Bayesian classifier [Labute 1999, Gao 1999].

Bayesian regularized artificial neural networks were employed [Burden 1999] to derive QSAR models, and perform better than regression methods that are not able to model nonlinearities in the model.

Artificial Neural Networks (ANNs) have been used to distinguish drug-like and non-drug-like molecules using a substructural analysis [Jain 1998]. So and Karplus [So 1997] used electrostatic and steric properties at grid points for feeding a genetic artificial neural network in order to develop a QSAR model.

Support Vector Machines (SVMs) attempt to learn the maximum separating boundary compared to Neural Networks which do not optimise the decision boundary if the prediction performance does not change. Compared to C5.0 decision trees, multi-layer perceptrons and other neural networks [Burbidge 2001], SVMs need less training time and achieve slightly better prediction performance. Using SVMs, Warmuth et al

[Warmuth 2003] implemented a concept of active learning. For other applications of Support Vector Machines in chemometrics, see [Czerminski 2001, Hearst 1998].

King [King 1992, King 1995] and Srinivasan and King [Srinivasan 1999] applied inductive logic programming (ILP) to the field of activity of molecules.

A general overview of structural representation, molecular similarity and virtual screening can be found in [Artymiuk 1992, Bures 1994, Livingstone 2000, Sheridan 2002, Bajorath 2002, Bajorath 2001, Willett 1992, Willett 1995, Willett 1998, Willett 2000, Walters 1998, Doucet 1996, Gillett 1998b, Good 1998]. An attempt to characterize molecular similarity methods is given in [Johnson 1991].

The method we present in the following section 3 is based on the 2-dimensional structure of molecules. It is derived from the connectivity table, thus not dependent on conformation and translationally as well as rotationally and conformationally invariant. Using the classification given above, it belongs to the group of subshape-based molecular representations, combined with a machine learning approach in the form of a Naïve Bayesian Classifier for classification of structures.

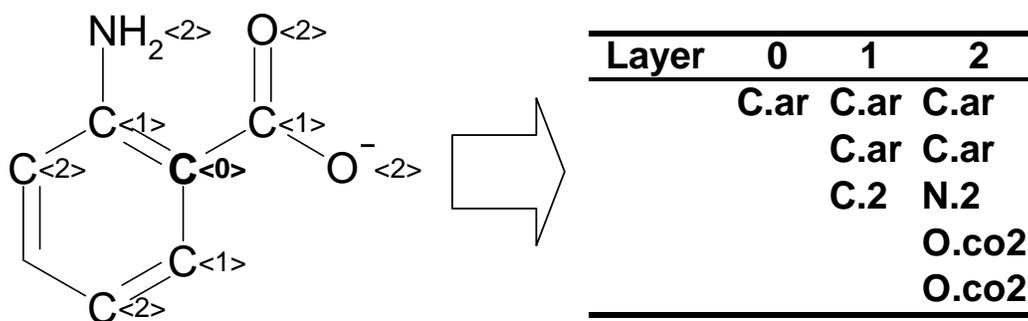
### 3. EXPERIMENTAL DETAILS

#### a) Descriptor Generation / Molecular Representation

We use atom environments [Xing 2002] as a molecular representation. Atom environments are similar to Signature Molecular Descriptors [Faulon 2003a; Faulon 2003b; Visco Jr. 2002; Faulon 1994]. They are translationally and rotationally invariant. Furthermore they do not depend on a particular conformation as they are calculated from the connectivity table. This makes generating atom environments less difficult compared to alignment-dependent approaches. Another benefit with atom environments is that they are easily interpretable, as they resemble the chemical concept of functional groups.

We calculated atom environments in a two-step procedure (see Figure 1):

1. Sybyl atom types [Clark 1989] are employed for the derivation of the environments. These are force-field atom types, which implicitly include molecular properties such as geometry. An individual atom fingerprint is calculated for every atom in the molecule. This calculation is performed using distances from 0 up to n bonds and keeping count of the occurrences of the atom types. The maximum distance n for descriptor generation has been varied from 1 to 3 for parameter optimization; details are given in section 3e.
2. A count vector is constructed with the vector elements being counts of atom types at a given distance from the central atom. Every atom is described by exactly one count vector resulting in molecular atom environment fingerprints in which the number of atoms in a given molecule equals the number of count vector entries in the fingerprint.



**Figure 1.** Illustration of descriptor generation step, applied to an aromatic carbon atom. The distances (“layers”) from the central atom are given in brackets. In the first step, Sybyl mol2 atom types are assigned to all atoms in the molecule. In the second step, count vectors from the central atom (here C<0>) up to a given distance (here two bonds from the central atom apart) are constructed. Molecular Atom Environment fingerprints are then binary presence/absence indicators of count vectors of atom types.

### b) Feature Selection

The information content of individual atom environments was computed using the information gain measure of Quinlan [Quinlan 1986, Glen 1992]. For a particular descriptor, higher information gain is related to better separation between active and inactive structures, for example.

The information gain,  $I$ , can be given by

$$I = S - \sum_v \frac{|S_v|}{|S|} S_v$$

Where

$$S = -\sum p \log_2 p$$

$S$  is the information entropy;  $|S|$  is the total number of data sets;  $S_v$  is the information entropy in data subset  $v$ ; and  $|S_v|$  is the number of data sets in subset  $v$ .

In each run the number of selected features was varied between 10 and 100.

### c) Classification

A Naïve Bayesian Classifier [Mitchell 1997] was employed as a classification tool. The Naïve Bayesian Classifier provides a simple yet surprisingly accurate machine-learning tool [Mitchell 1997]. Trained with a given data set which consists of known feature vectors ( $F$ ) and their associated known classes ( $CL$ ), the Naïve Bayesian Classifier predicts the class that a new feature vector belongs to as the one with the highest probability of  $P(CL_m | F)$  which is given by

$$P(CL_m | F) = \frac{P(CL_m)P(F | CL_m)}{P(F)} \quad (1)$$

Where

$P(CL_m)$ : probability of class  $m$

$P(F)$ : feature vector probability and

$P(F|CL_m)$ : probability of  $F$  given  $CL_m$

$m$  : class.

In the Naïve Bayesian Classifier

$$P(F | CL_m) = \prod_i P(f_i | CL_m)$$

Where,  $f_i$  are the feature vector elements. Hence for  $CL_m$ , (1) becomes

$$P(CL_m | F) = \frac{P(CL_m) \prod_i P(f_i | CL_m)}{P(F)}.$$

In this work the data are classified into two classes (active and inactive, here referred to as 1 and 2 respectively). Therefore

$$P(CL_1 | F) = \frac{P(CL_1) \prod_i P(f_i | CL_1)}{P(F)}$$

and

$$P(CL_2 | F) = \frac{P(CL_2) \prod_i P(f_i | CL_2)}{P(F)}$$

$$\Rightarrow \frac{P(CL_1 | F)}{P(CL_2 | F)} = \frac{P(CL_1) \prod_i P(f_i | CL_1)}{P(CL_2) \prod_i P(f_i | CL_2)}$$

$$\Rightarrow \frac{P(CL_1 | F)}{P(CL_2 | F)} = \frac{P(CL_1)}{P(CL_2)} \prod_i \frac{P(f_i | CL_1)}{P(f_i | CL_2)} \quad (2).$$

We use this equation to do the classification i.e., all molecules are represented by their feature vectors  $F$  and the resulting ratios  $\frac{P(CL_1 | F)}{P(CL_2 | F)}$  are sorted in decreasing order.

Molecules with the highest probability ratios are most likely to belong to class 1 (here the class of active molecules). Molecules with the lowest values are most likely to belong to class 2 (the class of inactive molecules).

Note that the actual probability  $P(CL_1 | F)$  can be easily computed from  $\ln\left(\frac{P(CL_1 | F)}{P(CL_2 | F)}\right)$  based on the fact that  $P(CL_1 | F) + P(CL_2 | F) = 1$ .

#### d) Compilation of Dataset and Pre-processing

For evaluation of the algorithm, 957 ligands extracted from the MDDR database [MDL] were used [Briem 2000]. The set contains 49 5HT3 Receptor antagonists (from now on referred to as 5HT3), 40 Angiotensin Converting Enzyme inhibitors (ACE), 111 3-Hydroxy-3-Methyl-Glutaryl-Coenzyme A Reductase inhibitors (HMG), 134 Platelet Activating Factor antagonists (PAF) and 49 Thromboxane A2 antagonists (TXA2). An additional 547 compounds were selected randomly and did not (according to MDDR) belong to any of these activity classes.

Structures were downloaded in SDF format and converted to Sybyl mol2 format using OpenBabel [OPENBABEL] 1.100.1 with the `-d` option to delete hydrogen atoms and default mol2 atom typing. Atom environment fingerprints were then calculated directly from mol2 files.

### e) Calculations

Two separate validations of the method presented here were performed. In the first validation, cross-validation with random selection of query molecules was carried out to optimize the parameters related to descriptor generation and feature selection. A 20-fold cross validation study selecting randomly five query structures for query generation and calculation of the average enrichment factors of the first 20 and 50 molecules of the sorted library has been performed. The selection of five query structures is a realistic number if few ligands of a given target are known. In order to illustrate the influence of the number of structures chosen to generate the query on search performance, 20-fold random selection of 3, 5 and 10 structures has been performed, selecting 40 features in the feature selection step. An individual hit rate was calculated for each set of compounds based on the number of molecules within its ten nearest neighbours, which belong to the same activity class as the query compound. To create a query from multiple molecules, individual probabilities (relative frequencies) of features from a set of molecules are calculated and used in the feature selection step described in section 3b and the Naïve Bayesian Classifier described in section 3c. The maximum bond distance for generation of molecular descriptors,  $n$ , was varied from 1 to 3. In each run, the number of selected features was set to 10, 20, 30, 40, 50, 70 and 100, starting with the features associated with highest information gain. To examine the influence of very frequent and very rare features, this series of experiments has been repeated with a slight modification. Using identical settings for maximum bond distance and number of selected features, only features occurring at least three times, but not in more than  $\text{max}-3$  molecules (with  $\text{max}$  being the number of molecules within the positive data set) were selected. To do so, features were chosen starting with those possessing the highest information

gain as above, but skipping rare and frequent features as defined here until the preset number of features was selected. For the best performing feature selection, cumulative recall plots were calculated for all five datasets of active compounds.

In all calculations presented here, the inactive dataset containing all structures except those of the active class in each calculation was split in two subsets of equal size to create independent training and test sets. Each similarity calculation was carried out twice, using the active query and each of the two subsets and scoring the remaining active compounds and the inactive compounds not used to generate the model. The average score of the active structures from both runs was calculated. Both subsets of scored inactive structures and the set of active structures with associated average scores were concatenated to give the complete scored list of compounds used for further processing. As an example, for one validation run using a sample of the ACE inhibitor dataset we have drawn the query molecules, selected fragment features and highest scoring molecules.

In the second validation, for each of the 383 active compounds of the five classes of active compounds its ten nearest neighbours were calculated based on the similarity measure proposed in sections 3a – 3c. The maximum distance for descriptor calculation was set to 2 as it produced the best results in the first validation run as well as in additional validations which were performed. Identical values for selection of features as mentioned above have been applied. Exclusion of frequent and rare atom environments was not applied due to the use of single query molecules. An individual hit rate was calculated for each compound based on the number of molecules within its ten nearest neighbors, which belong to the same activity class as the query compound. Enrichment is observed when the hit rate among the nearest neighbors is higher than the fraction of the activity class under consideration in the whole data set. Enrichments have been averaged over all classes of active compounds, and the result was compared to that of other methods.

Note that the nearest neighbour protocol of Briem [Briem 2000] has been followed in this validation to make it easy to compare the performance of our algorithm with commonly used methods.

## 4. RESULTS

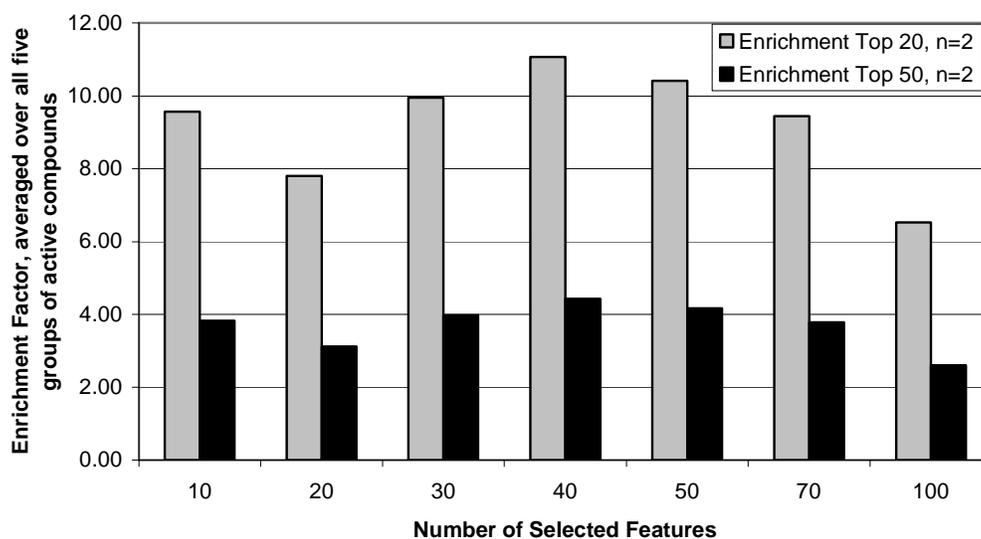
For the first validation, the influence of the maximum bond distance for creating the atom environment descriptor and the influence of the number of selected features on the average enrichment factor among the first 20 and the first 50 compounds of the ranked database are given in Table 1. Using atoms up to two bonds from the central atom for generating atom environment descriptors ( $n = 2$ ) produces best results with enrichment factors of between 11 and 6.5 in the first 20 compounds and between about 4 and 2 in the first 50 compounds. Using three layers for construction of the descriptor still gives enrichment of more than 3 in most cases of feature selection whereas using only the first layer adjacent to the central atom produces virtually no enrichment, independent of the method used for feature selection.

Number of Selected Features	Enrichment Top 20, n=1	Enrichment Top 50, n=1	Enrichment Top 20, n=2	Enrichment Top 50, n=2	Enrichment Top 20, n=3	Enrichment Top 50, n=3
10	0.00	0.00	9.56	3.82	3.50	1.40
20	0.00	0.00	7.80	3.12	3.50	1.40
30	0.00	0.00	9.95	3.98	3.50	1.40
40	0.00	0.00	11.06	4.42	3.50	1.40
50	0.00	0.00	10.42	4.17	2.92	1.17
70	0.00	0.00	9.45	3.78	3.11	1.24
100	0.00	0.00	6.52	2.61	0.19	0.08
10, $m > 2$ , $m < \max - 2$	0.00	0.00	7.33	2.93	3.50	1.40
20, $m > 2$ , $m < \max - 2$	0.00	0.00	8.58	3.43	3.50	1.40
30, $m > 2$ , $m < \max - 2$	0.00	0.00	9.20	3.68	3.50	1.40
40, $m > 2$ , $m < \max - 2$	0.00	0.00	10.28	4.11	3.50	1.40
50, $m > 2$ , $m < \max - 2$	0.00	0.00	10.13	4.05	3.50	1.40
70, $m > 2$ , $m < \max - 2$	0.07	0.03	8.99	3.59	3.50	1.40
100, $m > 2$ , $m < \max - 2$	0.07	0.03	4.89	1.96	0.19	0.08

**Table 1.** Enrichment factor averaged over all five classes of active compounds upon varying the number of selected features and the maximum depth,  $n$ , used to create the atom environment descriptor. A fixed number of features were selected and rare and frequent fragments were excluded. This is denoted by  $m > 2$ ,  $m < \max - 2$ , meaning that

features had to occur at least three times, but at most as often as the total number of active molecules (max) minus three times. In situations where very low enrichment factors were obtained many molecules were assigned identical scores, thus producing artefacts (enrichment factors of 0) in this table.

A visualization of enrichment factors, which depend on the number of selected features, is given in Figure 2. In this case, the bond level for descriptor generation,  $n$ , has been set to  $n = 2$  because it performed best as shown in Table 1. Exclusion of frequent and rare features does not perform as well as selection of a fixed number of features, and it is not shown in the figure.



**Figure 2.** Enrichment factor, averaged over all five groups of active compounds, using atoms up to 2 bonds apart from the central atom to construct the atom environment descriptor and a variable number of selected features for classification.

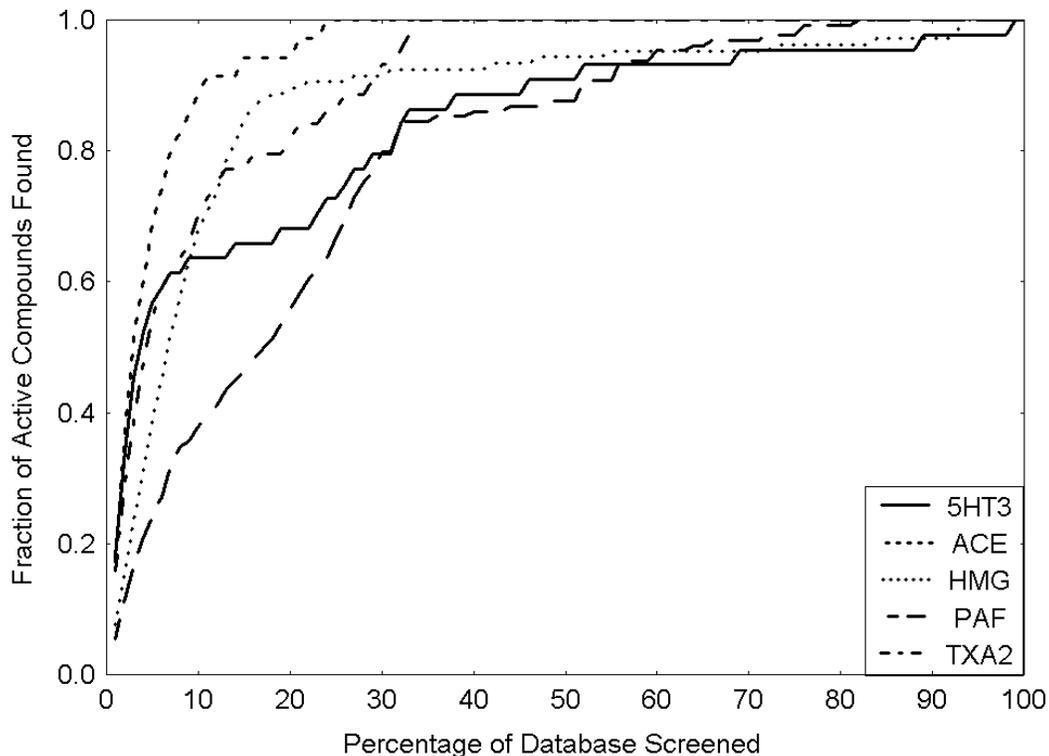
We have found that feature selection has its optimum at a selection of 40 features with respect to enrichment factors observed among the first 20 and among the first 50 highest-scoring structures of the sorted library. If fewer or more features are selected, performance of the algorithm continuously decreases.

The influence of the number of structures chosen to generate the query on search performance is shown in Table 2. Results using single structures to generate the active query are presented later and are included here for completeness. In every case except one (going from 5 to 10 query structures using ACE inhibitors), performance improves as the number of compounds used for query generation increases. The average deviation in performance between different sets of query compounds decreases if the size of the query data set is increased. Again, the only exception is if the number of ACE inhibitors used to generate the query is increased from 5 to 10 structures.

No. of query structures	5HT3	Std.-Dev.	ACE	Std.-Dev.	HMG	Std.-Dev.	PAF	Std.-Dev.	TXA2	Std.-Dev.	Mean	Std.-Dev.
<b>1</b>	5.65	4.26	6.40	2.96	7.90	2.75	7.15	2.25	6.40	3.27	6.70	3.10
<b>3</b>	8.55	1.73	6.70	2.64	9.30	0.92	9.15	1.57	8.30	1.13	8.40	1.60
<b>5</b>	9.25	1.02	9.10	0.64	9.50	0.83	9.15	0.82	8.15	1.04	9.03	0.87
<b>10</b>	9.30	1.03	8.80	1.51	9.70	0.57	9.25	0.72	8.95	0.76	9.20	0.92

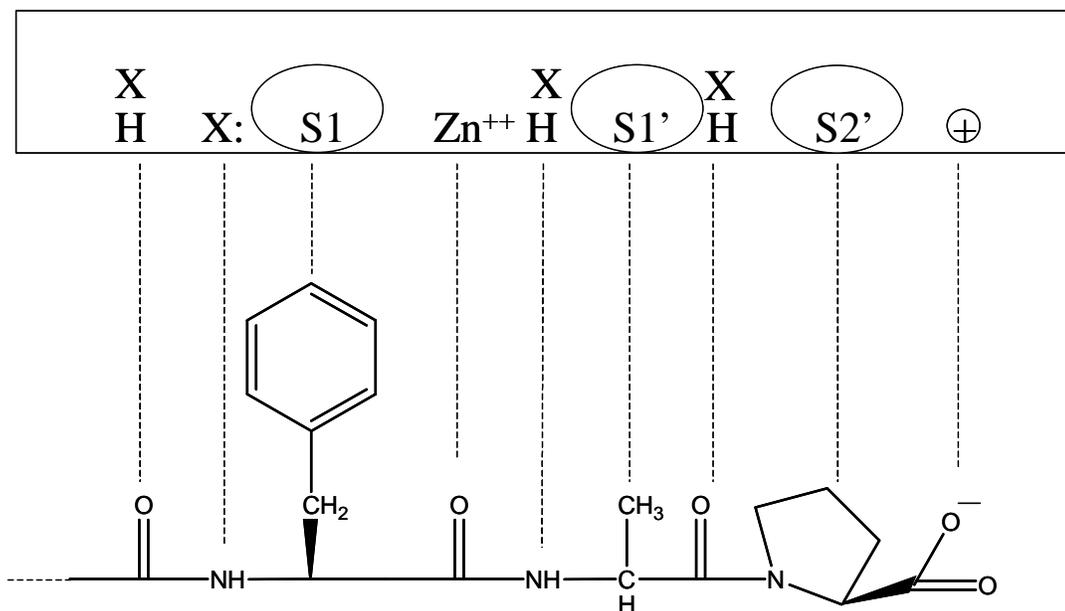
**Table 2.** Average hit rates among the ten nearest neighbors in a cross validation study. Shown here are the hit rates and the standard deviations among different data set sizes used to generate the query.

For the best performing method using 40 features with the highest information gain, cumulative recall plots are given in Figure 3. These plots were calculated using the 20-fold random selection of five queries for ranking of the library and screening for the remaining active compounds. The five datasets can be classified into two groups: The 5HT3, HMG and PAF datasets belong to one group as some of their active molecules are found only after evaluating half of the sorted library. ACE and TXA2 belong to the second group with all active molecules found well within the first 40% of the sorted library.



**Figure 3.** Cumulative recall plot of all five datasets, using atoms up to two bonds apart from the central atom for descriptor generation and 40 features associated with the highest information gain for classification.

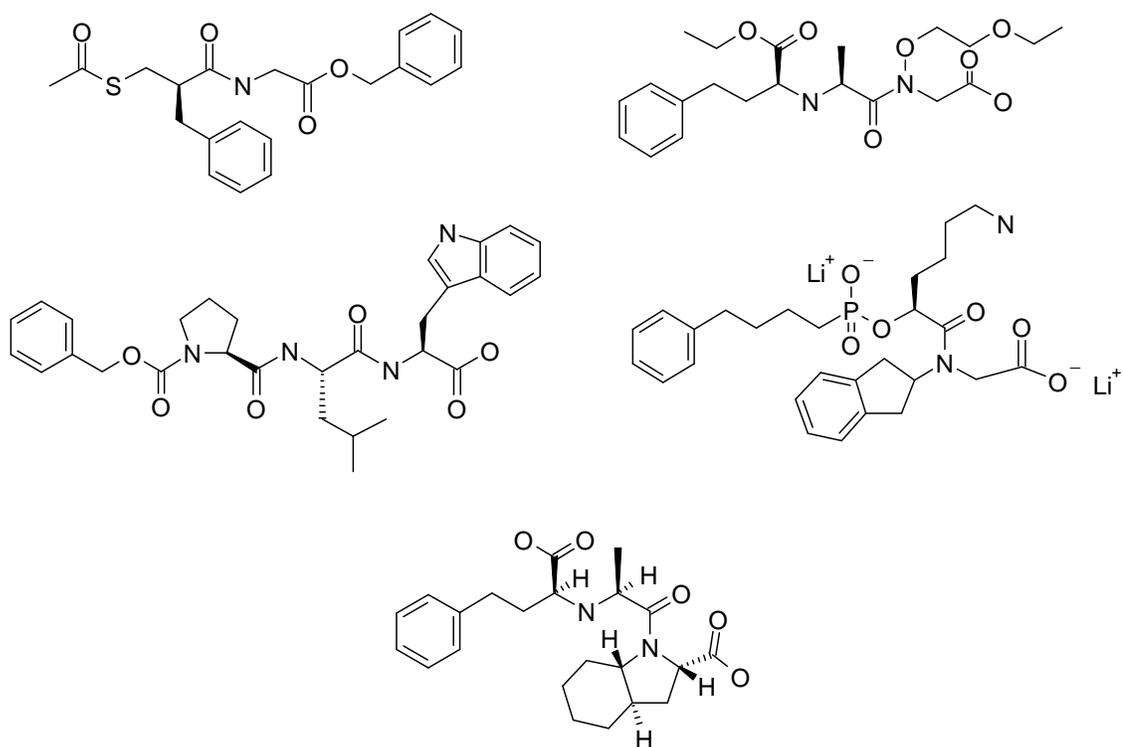
In order to gain an insight into the algorithm, query molecules, selected features and the highest scoring structures of the sorted library have been plotted for a sample run using angiotensin converting enzyme inhibitors (ACE inhibitors). The design of ACE inhibitors originally followed the hypothesis that ACE had binding site homology with carboxypeptidase-A [Cushman 1977]. A number of interaction sites were proposed based on analogue design, shown in Figure 4.



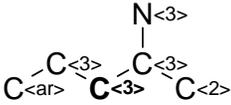
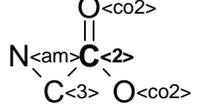
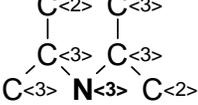
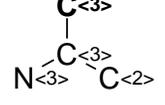
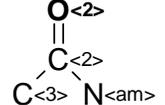
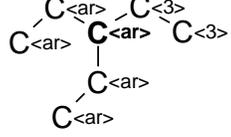
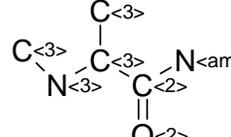
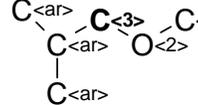
**Figure 4.** Snake venom peptide analogue with putative binding motif to angiotensin used in early compound design [Cushman 1977].

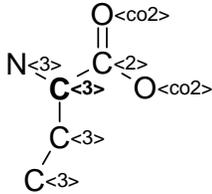
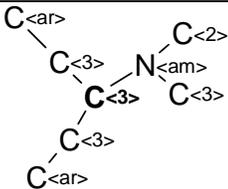
A recent crystallographic study of an ACE inhibitor, lisinopril (N2-[(S)-1-carboxy-3-phenylpropyl]-L-lysyl-L-proline), has revealed the binding site interactions in some detail [Natesh 2003]. Much of the originally deduced binding site topology is seen in the crystal structure with some notable differences such as the absence of the C-terminal carboxylate arginine interaction. The selection of features associated with a significant information gain in separating the classes of ACE and non-ACE inhibitors can be compared with the crystallographically determined binding motif. It may be expected that those interactions that are seen crystallographically may also emerge from the analysis of the analogues as being important.

The 5 molecules used to construct the query are shown in Figure 5, the 10 selected features giving the highest information gain are given in Table 3 and the 10 highest ranked structures from the sorted library are shown in Table 4.

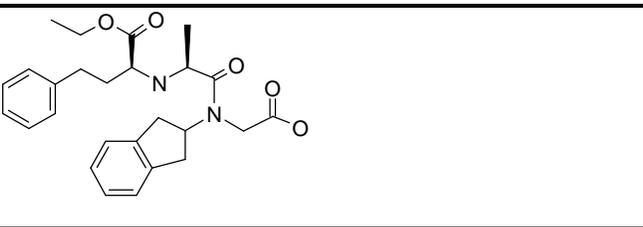
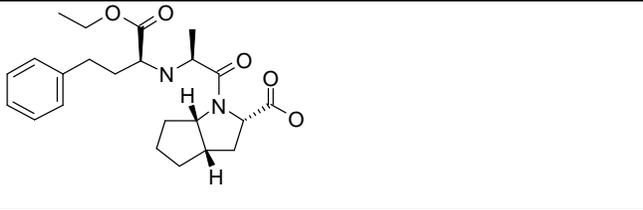
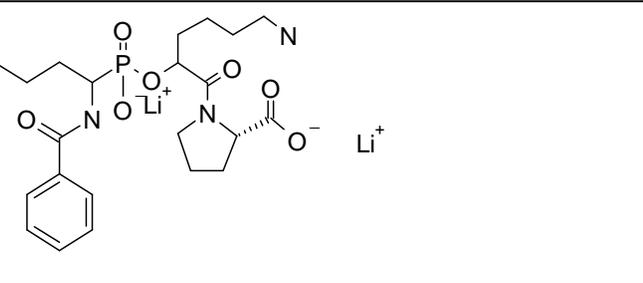
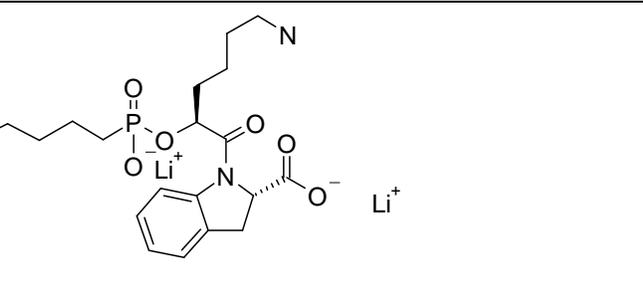
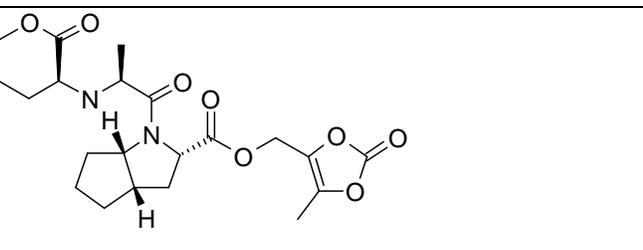


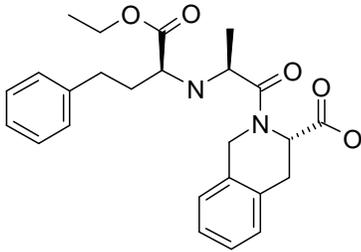
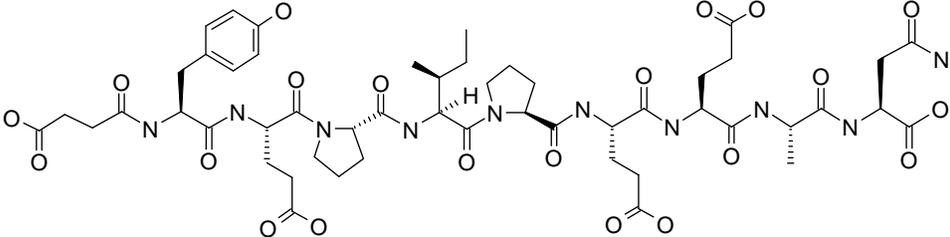
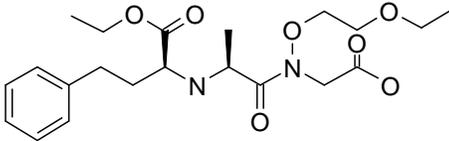
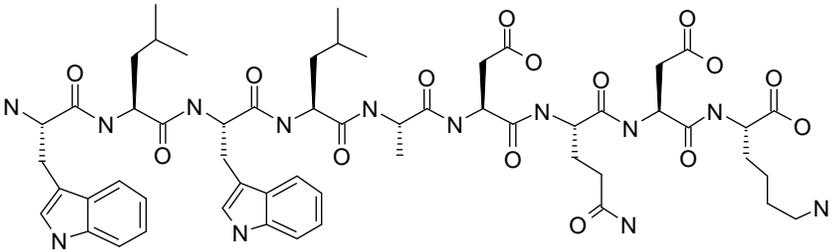
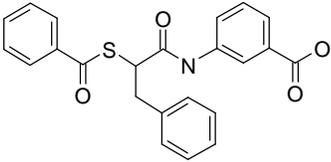
**Figure 5.** Five active molecules from the dataset of ACE inhibitors, used to construct the query and perform feature selection.

Selected Feature	Information Gain Associated with this Feature	Putative Interaction Site on ACE
	0.0171	S2
	0.0141	Zn <sup>++</sup>
	0.0127	S2
	0.0118	S1
	0.0114	XH/Zn <sup>++</sup>
	0.0104	S1
	0.0096	S2/+
	0.0086	S1

	0.0083	S2/+
	0.0083	S2

**Table 3.** Ten features associated with the highest information gains from a sample run using 5 inhibitors from the ACE dataset.

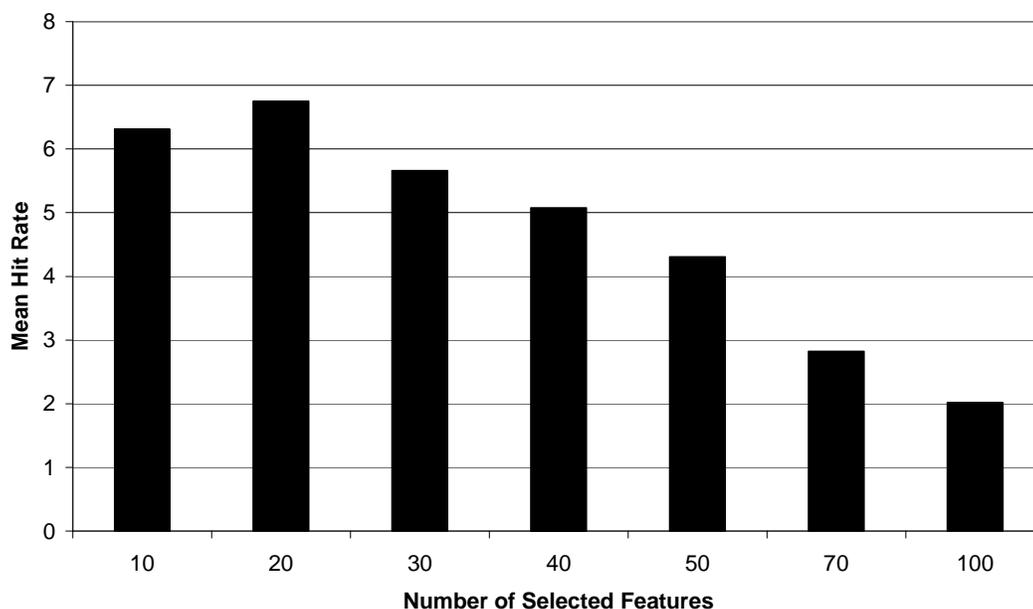
Rank number / Activity	Structure
1 / Active	
2 / Active	
3 / Inactive	
4 / Active	
5 / Active	

6 / Active	
7 / Inactive	
8 / Active	
9 / Inactive	
10 / Active	

**Table 4.** Top 10 ranking molecules of the sorted library. Out of these, seven are active ACE inhibitors and three are inactive molecules in this respect.

The selected features, given in Table 3, possess carbon, nitrogen as well as oxygen atoms as central atoms. Some analysis of the selected features with respect to the experimentally determined interaction of ligands within the ACE binding site is given in the discussion section.

In the second validation, the number of active molecules among the ten nearest neighbours of each individual active molecule from each dataset was calculated following the protocol of Briem and Lessel [Briem 2000]. Feature selection was performed selecting 10, 20, 30, 40, 50, 70 and 100 features. Hit rates, averaged over all five classes of active compounds are given in Figure 6.



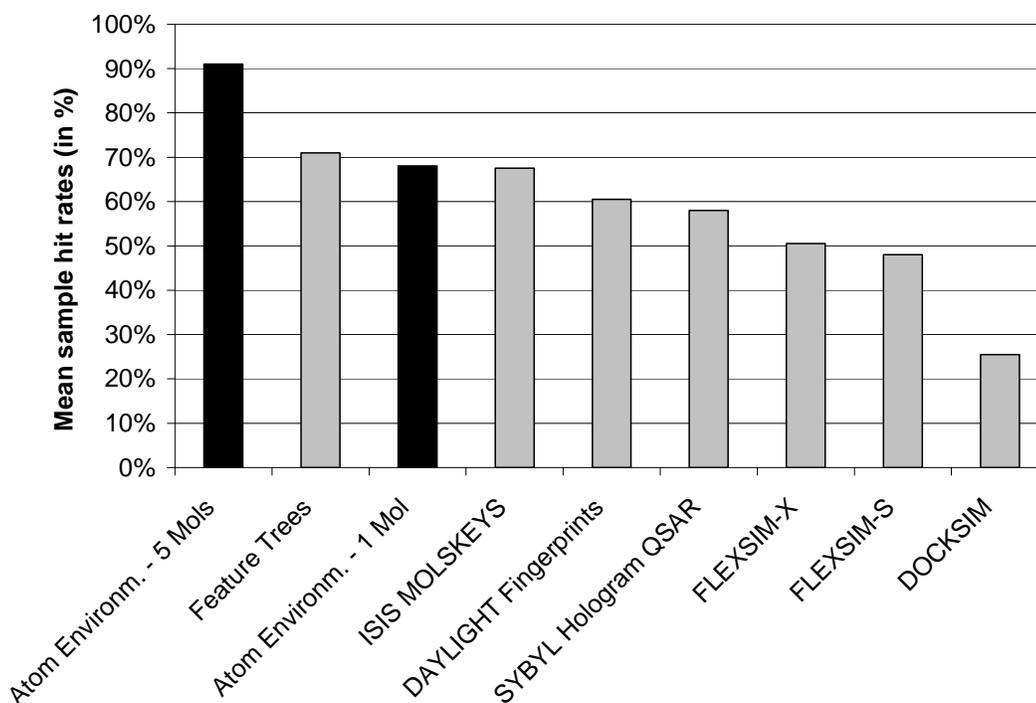
**Figure 6.** Mean hit rates among the ten nearest neighbours of the Briem data set, averaged over all five classes of active compounds, depending on the number of selected features.

An optimum can be seen at the selection of 20 features. If more or less features are used for classification, performance declines continuously. The individual hit rates for each group, using the best-performing selection of 20 features, are given in Table 5. The average number of active compounds among the top 10 ranked compounds varies from about 5.65 (5HT3) to about 7.90 (HMG), with an overall average of 6.70. These numbers, taking into account the variable number of active structures in each active subset result in enrichment factors between 5.14 (PAF) and 15.6 (ACE). The overall average enrichment factor calculates to 8.48, which is significantly higher than the value of 1 that would be achieved in a random selection.

Performance of the Atom Environment Approach, Selecting 20 Features						
Group of Active Compounds	5HT3	ACE	HMG	PAF	TXA2	Overall
Expected Hit Rate	0.50	0.41	1.15	1.39	0.50	0.79
Average Number of Active Compounds Among Top 10 Ranked Compounds						
	5.65	6.40	7.90	7.15	6.40	6.70
Enrichment Factor	11.0	15.6	6.87	5.14	12.8	8.48

**Table 5.** Performance of the Atom Environment Approach by measuring mean sample hit rates of the ten top-scored compounds in the sorted hit list. Feature selection was performed selecting 20 features associated with the highest information gain.

The nearest neighbour protocol of Briem [Briem 2000] has been followed in this validation to enable ease of comparison of the algorithm performance with established methods. The methods used for comparison are Feature Trees [Rarey 1998], ISIS MOLSKKEYS [ISIS], Daylight Fingerprints [DAYLIGHT], SYBYL Hologram QSAR Fingerprints [SYBYL] and FLEXSIM-X [Lessel 2000], FLEXSIM-S [Lemmen 1998] and DOCKSIM [Briem 1996] virtual affinity fingerprints. Feature Trees represent molecules as trees (acyclic graphs), which are subsequently matched for comparison. In current versions, FlexX interaction profile and Van-der-Waals radii have been used as descriptors and a size-weighted ratio of fragments is used to calculate a similarity index. ISIS MOLSKKEYS use 166 predefined two-dimensional fragments for describing a structure. Daylight Fingerprints are algorithmically generated and describe atom paths of variable length: they are commonly folded and a 1024 bits long bit string is used. Hologram QSAR is an extension of 2D fingerprints and additionally includes branched and cyclic fragments as well as stereochemical information. For all 2D and 3D descriptors, Euclidean distances were calculated for each possible combination of test ligands. The performance of the algorithm presented here compared to established methods is shown in Figure 7.



**Figure 7.** Mean sample hit rates of the Atom Environment approach (black), in comparison to the methods applied by Briem (light grey). The performance of the Atom Environment approach is shown using single queries and randomly selected subsets of five query molecules.

Shown here are mean sample hit rates as averaged over all five classes of active compounds. Using one query structure, this method outperforms all three virtual affinity fingerprint algorithms as well as two of the two-dimensional methods, Daylight Fingerprints and SYBYL Hologram QSAR Fingerprints. It performs as well as ISIS MOLSKEYS fingerprints and is only (marginally) outperformed by the Feature Tree approach. The top three methods are of comparable performance, however the atom environments approach additionally deduces those fragments having the greatest influence on similarity and is significantly faster than Feature Trees and therefore of utility in searching larger databases.

Using five query structures, the Atom Environment approach achieves a mean sample hit rate of greater than 90%.

The computation of molecular fingerprints was implemented in C programming language and was able to process about 1000 molecules per second on a Pentium III-1GHz workstation. Feature selection and scoring was implemented in Perl and was able to evaluate one molecule against the 956 remaining compounds of the dataset in one second, using identical hardware.

## 5. DISCUSSION

The first series of runs was performed to optimize parameters of the algorithm for typical database screenings where several active compounds are known. As Table 1 shows, the algorithm only gives sensible results when the atom environment descriptor is constructed using atoms up to two bonds apart from the central atom. If less than two bonds are considered, atom environments are ambiguous and do not capture enough information about the atom environment. If more than two bonds are considered, they tend to become unique so no generalization capability is acquired. This result is in agreement with the results found by Faulon et al. [Faulon 2003; Faulon 2003b]. Optimum performance is found with the selection of 40 features. This is the result for queries derived from five query structures and applies across the five different sets of active molecules used. Fewer features do not allow the classification of each molecule reliably (by recognizing a certain number of its atom environments) and more features appear to introduce noise into the system thus reducing its classification ability.

The performance of the algorithm generally increases if more and more structures are used to generate the query (Table 2), as well as the standard deviation in performance between different sets of query structures decreases. Using five query structures, Atom Environments outperforms other methods by a large margin (Figure 7), giving mean sample hit rates of about 90%. These hit rates are not directly comparable, because information from multiple structures is used to formulate the query. Nonetheless, it shows that the algorithm is capable of handling information from multiple molecules reliably. For real-world applications, it appears that all active molecules across the range of structural diversity could be used in order to train the classifier used in this method.

The five datasets used can be classified into two groups. In one group, comprising the 5HT3, HMG and PAF datasets, hits are still found among lower ranked molecules (Figure 3). Apparently, there are molecules in these groups of datasets which do not possess close analogues in the training and the test set. In the other group, comprising ACE and TXA2, all active molecules are easily found in the first half of the focused

library. The molecules in these classes of active compounds seem to be more similar to each other.

Overall the selection of fragments of ACE inhibitors seems consistent with the binding information deduced crystallographically [Natesh 2003]. The five fragments associated with the highest information gain given in Table 3 correspond to the binding motif of enalapril and captopril including the zinc binding site and the S2 and S1 sites in the top rank. Among the 10 highest scoring molecules of the sorted library listed in Table 4, seven are known active ACE inhibitors while three are not tested with respect to ACE inhibitor activity. The inactives (which of course, may be active – the data on these molecules in MDDR do not include ACE assay results) are peptidic, larger than small molecule analogues and contain many peptidic environments common to the natural substrates. Elimination of such peptidic moieties would give (in this case) an ideal result. A penalty factor for molecules larger than the probe molecules (a scoring relative to size) could be used.

When calculating the hit rate of the ten nearest neighbours of each individual active molecule (i.e. using one molecule to retrieve its neighbours from the remaining database), an optimum in classification is obtained if 20 features are selected (Figure 6). This is a different result from that observed in those runs where five molecules are used to derive the query. In those cases, the optimum feature number was seen to increase to 40 features. In single molecule queries, one active molecule containing only a small number of atom environments is used. An increase in the number of features thus exceeds the number of environments present in many molecules. Therefore, there is no gain in including additional features.

As described in Table 5, enrichment factors have been found to be between 5.14 (PAF) and 15.6 (ACE). The overall average enrichment factor is 8.48, showing general validity of the approach.

The method presented here and all top-performing algorithms it is compared to are two-dimensional approaches. Two-dimensional similarity searching algorithms often lead to surprisingly good results. However, one has to be careful that this is not simply due to the libraries used which often contain analogue structures. Analogue design is commonly successful in finding new active molecules and analogue

molecules often contain identical substructures. Two-dimensional algorithms, which are based on connectivity tables, easily detect these identical substructures. This is a general problem of compiling databases for evaluating database retrieval performance and affects all of the algorithms employed in this work.

Using single queries, Feature Trees, Atom Environments and ISIS MOLSKEYS perform considerably better than Daylight Fingerprints and Hologram QSAR on the test sets. The latter group has in common that it includes information in addition to local subgraph features, whereas the former group only uses local information. This is the case because Feature Trees are commonly repeatedly cut before matching, ISIS MOLKEYS use predefined fragments and Atom Environments only consider an atom and its neighbours at a maximum of two bonds apart. Restricting molecular representation to local information might therefore be a useful feature.

In addition, ISIS MOLSKEYS and Atom Environments employ feature selection. ISIS MOLSKEYS considers only fragments occurring in a library whereas in the case of Atom Environments, fragments are explicitly selected. Daylight Fingerprints, on the other hand, consider every atom path in a certain distance range and then fold the information to give uniform length descriptors. The lack of feature selection or the hashing and folding process seems to worsen the performance of this type of descriptor.

All three virtual affinity fingerprint methods perform worse than any of the two-dimensional methods when applied to the test datasets. Virtual affinity methods consider the three-dimensional structure of the ligand and also take the structure of the receptor into account. Probably because of currently used strategies of library design, as mentioned above, the performance of three-dimensional virtual affinity fingerprint methods is generally seen to be lower than the performance of two-dimensional methods. Nonetheless, it is reported that three-dimensional similarity measures are able to detect similarities which two-dimensional methods are unable to pick up [Briem 2000]. This would be true in particular in the case of conformationally labile molecules which can achieve pharmacophoric patterns that are important for

activity or stereochemically important combinations which are not encoded in the 2D representation.

Briem [Briem 2000] gives more details about variations in performance among virtual affinity fingerprint based techniques.

## 6. CONCLUSIONS

In this paper we introduced the combination of atom environments, information-gain based feature selection and a Naïve Bayesian Classifier to describe the similarity of molecules. On average, our algorithm achieved an enrichment factor of about 8 when calculating the ten nearest neighbours of five datasets containing active structures. In addition to this encouraging result, the algorithm was compared to several two- and three-dimensional methods. Using single queries, it performs as well as the best commonly used 2D algorithms while outperforming all 3D methods. Using multiple queries, close-to-ideal hit rates are obtained. The technique described in this paper can also be useful in identifying key functional groups in active molecules and is computationally efficient.

## 7. FUTURE WORK

We will explore further all three steps of similarity searching, i.e. description of molecules, feature selection and learning of the model. We shall assign properties instead of Sybyl mol2 types to focus on similar chemical properties instead of identical atom types. Right now, we are only using exact matching, although fuzzy matching (involving e.g. decay functions on comparison of atom environments) might perform better (in fact, preliminary analyses are very promising). It is also of interest to deduce whether differences in performance are a result of the descriptor used (e.g. atom environments vs. ISIS MOLSKEYS) or of different similarity metrics (Bayes Classifier vs. Bit String Similarity coefficients).

## REFERENCES

- [Aitchison 1976] Aitchison, J.; Aitken, C.G.G. Multivariate binary discrimination by the kernel method. *Biometrika* **1976**, 68, 413 – 420.
- [Artymiuk 1992] Artymiuk, P.J.; Bath, P.A.; Grindley, H.M.; Pepperrell, C.A.; Poirrette, A.R.; Rice, D.W.; Thorner, D.A.; Wild, D.J.; Willett, P.; Allen, F.H.; Taylor, R. Similarity Searching in Databases of 3-Dimensional Molecules and Macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 617 – 630.
- [Bajorath 2001] Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 233 – 245.
- [Bajorath 2002] Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nature Drug Discov.* **2002**, 1, 882 – 894.
- [Balaban 1982] Balaban, A.T. Highly discriminatory distance-based topological index. *Chem. Phys. Lett* **1982**, 89, 399 – 404.
- [Balaban 1995] Balaban, A.T. Chemical Graphs: Looking back and Glimpsing Ahead *J. Chem. Inf. Comput. Sci.* **1995**, 35, 339 – 350.
- [Barnard 1993] Barnard, J.M. Clustering of Chemical Structures on the Basis of 2-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1993**, 32, 644 – 649.
- [Bath 1994] Bath, P.A.; Poirrette, A.R.; Willett, P.; Allen, F.H. Similarity Searching in Files of 3-Dimensional Chemical Structures – Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 141 – 147.
- [Baumann 1999] Baumann, K. Uniform-length molecular descriptor for quantitative structure-property (QSPR), quantitative structure-activity (QSAR), classification studies and similarity search. *Trends Anal. Chem.* **1999**, 18, 36 – 46.
- [Bemis 1992] Bemis, G.W.; Kuntz, I.D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comp.-Aided Mol. Des.* **1992**, 6, 607 – 628.
- [Bender 2004] Bender, A.; Mussa, H.Y.; Glen, R.C.; Reiling, S. Molecular Similarity Searching using Atom Environments, Information-Based Feature Selection and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.*, accepted for publication (2003).
- [Briem 1996] Briem, H.; Kuntz, I.D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, 39, 3401 – 3408.
- [Briem 2000] Briem, H.; Lessel, U. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug. Discov. Des.* **2000**, 20, 231 – 244.
- [Burbidge 2001] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pattern Recognition. *Comput. Chem.* **2001**, 26, 5 – 14.
- [Burden 1999] Burden, F.R.; Winkler, D.A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, 42, 3183 – 3187.
- [Bures 1994] Bures, M.G.; Martin, Y.C.; Willett, P. Searching Techniques for Databases of 3-Dimensional Chemical Structures. *Topics Stereochem.* **1994**, 21, 467 – 511.
- [Carbó 1980] Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quantum Chem.* **1980**, 17, 1185 – 1189.
- [Carbo 1992] Carbo, R.; Calabuig, B. Quantum Similarity Measures, Molecular Cloud Descriptors and Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 600 – 606.
- [Carbo-Dorca 1998] Carbo-Dorca, R.; Besalu, E. A general survey of Molecular Quantum Similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, 451, 11 – 23.
- [Clark 1989] Clark, R.D.; Cramer, R.D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comp. Chem.* **1989**, 10, 982-1012.

- [Cone 1977] Cone, M.M.; Venkataraghavan, R.; McLafferty, F.W. Molecular Structure Computer Program for the identification of maximal common substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668 – 7671.
- [Cramer 1974] Cramer, R.D.; Redl, G.; Berkoff, C.E. Substructural analysis: A novel approach to the problem of drug design. *J. Med. Chem.* **1974**, *17*, 533 – 535.
- [Cramer 1988] Cramer, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular-Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959 – 5967.
- [Cushman 1977] Cushman, D.W.; Cheung, H.S.; Sabo, E. F.; Ondetti, M. A. Design of potent competitive inhibitors of angiotensin-converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. *Biochemistry* 1977, *16*, 5484-5491.
- [Czerminski 2001] Czerminski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. *Quant. Struct.-Act Rel.* **2001**, *20*, 227 – 240.
- [DAYLIGHT] DAYLIGHT, Version 4.62, DAYLIGHT Inc., Mission Viejo, California, USA.
- [DiMasi 2003] DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003**, *835*, 1–35.
- [Dixon 1998] Dixon, E.L.; Villar, H.O. Bioactive Screening Library Selection via Affinity Fingerprints. *J. Chem. Inf. Comput. Sci.* **1998**, *31*, 722 – 729.
- [Dixon 1999] Dixon, S.L.; Koehler, R.T. The hidden component of size in two-dimensional fragment descriptors: Side effects on sampling in bioactive libraries. *J. Med. Chem.* **1999**, *42*, 2887 – 2900.
- [Dixon 2001] Dixon, S.L.; Merz, K.M. One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation. *J. Med. Chem.* **2001**, *44*, 3795 – 3809.
- [Doucet 1996] Doucet, J.-P. In: *Computer-Aided Molecule Design: Theory and Applications*, Springer **1996**, 328 – 362.
- [Downs 1994] Downs, G.M.; Willett, P.; Fisanick, W. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094-1102.
- [Duca 2001] Duca, J.S.; Hopfinger, A.J. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1357 – 1387.
- [Estrada 2001] Estrada, E; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8*, 1573-1588.
- [Europarl2003]  
[http://wwwdb.europarl.eu.int/oeil/oeil\\_ViewDNL.ProcedureView?lang=2&procid=4179](http://wwwdb.europarl.eu.int/oeil/oeil_ViewDNL.ProcedureView?lang=2&procid=4179)
- [Faulon 1994] Faulon, J.L. Stochastic generator of chemical structure: 1. Application to the structure elucidation of large molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1204-1218.
- [Faulon 2003a] Faulon, J.L.; Visco Jr., D.P.; Pophale, R.S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707-720.
- [Faulon 2003b] Faulon, J.L.; Churchwell, C.J.; Visco Jr., D.P. The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721-734.
- [Fligner 2002] Fligner, M.A.; Verducci, J.S.; Blower, P.E. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110 – 119.
- [Free 1964] Free, S.M.; Wilson, J.W. "A Mathematical Contribution to Structure Activity Studies. *J. Med. Chem.* **1964**, *7*, 395 – 399.
- [Gaillard 1994] Gaillard, P; Carrupt, P.A.; Testa, B.; Boudon, A. Molecular lipophilic potential, a tool in 3D QSAR: methods and applications. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 83 – 96.
- [Gao 1999] Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure-activity relationship (qsar) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164, 168.

- [Ghuloum 1999] Ghuloum, A.M; Sage, C.R.; Jain, A.J. Molecular hashkeys: A novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J. Med. Chem.* **1999**, 42, 1739 – 1748.
- [Gillet 1998] Gillet, V.J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 165 – 179.
- [Gillett 1998b] Gillet, V.J.; Wild, D.J.; Willett, P.; Bradshaw, J. Similarity and Dissimilarity Methods for processing chemical structure databases. *The Computer Journal* **1998**, 41, 547 – 558.
- [Ginn 1997] Ginn, C.M.R; Turner, D.B.; Willett, P.; Ferguson, A.M.; Heritage, T.W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 23 – 37.
- [Glen 1992] Glen, R.C.; A-Razzak, M. Applications of Rule-Induction in the Derivation of quantitative structure-activity relationships. *J. Comp. Aid. Mol. Des.*, **1992**, 6, 349-383.
- [Godden 2000] Godden, J.W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 163 – 166.
- [Good 1992] Good, A.C.; Hodgkin, E.E.; Richards W.G. Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 188 – 191.
- [Good 1993] Good, A.C.; Richards, W.G. Rapid Evaluation of Shape Similarity using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 112 – 116.
- [Good 1995b] Good, A.C.; Ewing, T.J.; Gschwend, D.A.; Kuntz, I.D. New Molecular Shape Descriptors: Applications in Database Screening. *J. Comput.-Aided Mol. Des.* **1995**, 9, 1 – 12.
- [Good 1998] Good, A.C.; Richards, W.G. Explicit Calculation of 3D molecular similarity. *Perspect. Drug Discov. Des.* **1998**, 9-11, 321 – 338.
- [Grant 1995] Grant, J.A.; Pickup, B.T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, 99, 3503 – 3510.
- [Gund 1977] Gund, P. Three-dimensional pharmacophoric pattern searching. *Progr. Mol. Subcell. Biol.* **1977**, 5, 117 – 143.
- [Hagadone 1992] Hagadone, T.R. Molecular Substructure Searching: Efficient retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 515 – 521.
- [Hall 1995] Hall, L.H.; Kier, L.B. Electrotopological State Indexes for Atom Types – A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 1039 – 1045.
- [Harper 2001] Harper, G.; Bradshaw, J.; Gittins, J.C.; Green, D.V.S.; Leach, A.R. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1295 – 1300.
- [Hearst 1998] Hearst, M.A. Support vector machines. *IEEE Intelligent Systems* **1998**, 18-28.
- [Hodgkin 1987] Hodgkin, E.E.; Richards, G.W. Molecular Similarity based on electrostatic potential and electric field. *Int. J. Quant. Chem. Quant. Biol. Symp.* **1987**, 14, 105 – 110.
- [Holliday 2002] Holliday, J.D.; Hu, C-Y.; Willett, P. Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. *Combin. Chem. High-Thr. Scr.* **2002**, 5, 155-166.
- [Holliday 2003] Holliday, J.D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 819 – 828.
- [Hubalek 1982] Hubalek, Z. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol. Rev.* 1982, 57, 669-689.
- [Ihlenfeldt 1994] Ihlenfeldt, W.D.; Gasteiger, J. Hash Codes for the Identification and Classification of Molecular Structure Elements. *J. Comput. Chem.* **1994**, 15, 793 – 813.
- [ISIS] ISIS, Version 2.1.4, Molecular Design Ltd., San Leandro, USA.

- [Jain 1994] Jain, A.N.; Koile, K.; Chapman, D. Compass: Predicting biological activities from molecular surface properties. Performance comparison on a steroid benchmark. *J. Med. Chem.* **1994**, 37, 2315 – 2327.
- [Jain 1998] Jain, A.N.; Walters, P.W.; Murcko, M.A. Can we learn to distinguish between "Drug-like" and "Nondrug-Like" Molecules? *J. Med. Chem.* **1998**, 41, 3314 – 3324.
- [Jain 2000] Jain, A.N. Morphological similarity: A 3D Molecular Similarity Method: Correlation with Protein-Ligand Interactions. *J. Comput.-Aided Mol. Des.* **2000**, 14, 199 – 213.
- [Jakes 1987] Jakes, S.E.; Watts, N.; Willett, P.; Bawden, D.; Fisher, J.D. Pharmacophoric Pattern-Matching in Files of 3D Chemical Structures – Evaluation of Search Performance. *J. Mol. Graph.* **1987**, 5, 41 – 48.
- [Johnson 1991] *Concepts and Applications of Molecular Similarity*; Johnson, A.M.; Maggiora, G.M., Ed.; Wiley: New York, 1990.
- [Kauvar 1995] Kauvar, L.M.; Higgins, D.L.; Villar, H.O.; Sportsman, J.R.; Engquist-Goldstein, A.E.; Bukar, R.; Bauer, K.E.; Dilley, H.; Rocke, D.M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, 2, 107 – 118.
- [Kellogg 1996] Kellogg, G.E.; Kier, L.B.; Gaillard, P.; Hall, L.H. E-state fields: application to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, 10, 513 – 520.
- [King 1992] King, E.D.; Muggleton, S.; Lewis, R.A.; Sternberg, M.J.E. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Nat. Acad. Sci. USA* **1992**, 89, 11322-11326.
- [King 1995] <http://citeseer.nj.nec.com/king95relating.html>
- [Klebe 1994] Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, 37, 4130 – 4146.
- [Klebe 1998] Klebe, G. Comparative molecular similarity indices analysis: CoMSIA. *Perspect. Drug Discov. Des.* **1998**, 12, 87 – 104.
- [Labute 1999] Labute, P. Binary QSAR: a new method for the determination of quantitative structure-activity relationships. *Pac. Symp. Biocomp.* **1999**, 4, 444 – 455.
- [Leicester 1988] Leicester, S.E.; Finney, J.L.; Bywater, R.P. Description of molecular surface shape using Fourier descriptors. *J. Mol. Graph.* **1988**, 6, 104 – 108.
- [Lessel 2000] Lessel, U.F.; Briem, H. Flexsim-X: A Method for the Detection of Molecules with Similar Biological Activity. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 246-253.
- [Lipinski 1997] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and Computational Approaches to Estimating Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug. Del. Des.* **1997**, 23, 3-25.
- [Lipinski 2000] Lipinski, C.A. Drugs-like properties and the causes for poor solubility and poor permeability. *J. Pharmacol. Toxicol. Meth.* **2000**, 44, 235 – 249.
- [Livingstone 2000] Livingstone, D.J. The characterization of chemical structures using molecular properties. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 195 0 209.
- [Martin 1992] Martin, Y.C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, 35, 2145 – 2154.
- [Mason 1999] Mason, J.S.; Cheney, D.L. Ligand-Receptor 3-D Similarity Studies Using Multiple 4-Point Pharmacophores. *Pac. Symp. Biocomp.* **1999**, 4, 456 – 467.
- [Mason 2001] Mason, J.S.; Good, A.C.; Martin, E.J. 3D-Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, 7, 567-597.
- [MDL] MDL Drug Data Report; MDL ISIS/HOST software, MDL Information Systems, Inc.
- [Mitchell 1997] *Machine Learning*; Mitchell T.M.: McGraw-Hill: New York, 1997.

- [Mount 1999] Mount, J.; Ruppert, J.; Welch, W.; Jain, A.J. A Flexible surface-based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60–66.
- [Natesh 2003] Natesh, R.; Schwager, S.L.U.; Sturrock, E.D.; Acharya, K.R. Crystal structure of the human angiotensin-converting enzyme–lisinopril complex. *Nature* **2003**, *421*, 551–554.
- [Nilakantan 1993] Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A New Method for Rapid Characterization of Molecular Shape: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79–85.
- [OPENBABEL] OpenBabel, <http://openbabel.sourceforge.net/>.
- [Pastor 2000] Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRIND-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3243.
- [Pickett 1996] Pickett, S.D.; Mason, J.S.; McLay, I.M. Diversity Profiling and Design using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1223.
- [Quinlan 1986] Quinlan, J.R. Induction of Decision Trees. *Machine Learning* **1986**, *1*, 81-106.
- [Randic 1979] Randic, M.; Wilkins C.L. Graph theoretical approach to recognition of structural similarity in molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31–37.
- [Rarey 1998] Rarey, M.; Dixon, J.S. Feature Trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471-490.
- [Rouvray 1992] Rouvray, D.H. Definition and Role of Similarity Concepts in the Chemical and Physical Sciences. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 580–586.
- [Schoonjans 2001] Schoonjans, V.; Questier, F.; Guo, Q.; van der Heyden, Y.; Massart, D.L. Assessing molecular similarity/diversity of chemical structures by FT-IR spectroscopy. *J. Pharm. Biomed. Anal.* **2001**, *24*, 613–627.
- [Schoor 1996] Schuur, J.H.; Selzer, P.; Gasteiger, J. The coding of the Three-Dimensional Structure of Molecules by Molecular Transformations and its Applications of Structure-Spectra Calculations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.
- [Sheridan 1989] Sheridan, R.P.; Ramaswamy, N.; Rusinko III., A.; Bauman, N.; Haraki, K.S.; Venkataraghavan, R. 3D Search: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255–260.
- [Sheridan 1996] Sheridan, R.P.; Miller, M.D.; Underwood, D.J.; Kearsley, S.K. Chemical Similarity Using Geometric Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- [Sheridan 2002] Sheridan, R.P.; Kearsley, S.K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2002**, *7*, 903–911.
- [So 1997] So, S.-S.; Karplus, M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. *J. Med. Chem.* **1997**, *40*, 4347–4359.
- [Soltzberg 1976] Soltzberg, L.J.; Wilkins, C.L. Molecular Transforms: A Potential Tool for Structure-Activity Studies. *J. Am. Chem. Soc.* **1977**, *99*, 439–443.
- [Srinivasan 1999] Srinivasan, A.; King, R.D. Feature construction with inductive logic programming: a study of quantitative predictions of biological activity aided by structural attributes. *Knowledge Discov. Data Mining J.* **1999**, *3*, 37–57.
- [Stanton 1990] Stanton, D.; Jurs, P. Development and Use of Charged Partial Surface-Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329.
- [Stiefl 2003] Stiefl, N.; Baumann K. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure-Activity Relationship Technique. *J. Chem. Inf. Comput. Sci.* **2003**, *46*, 1390–1407.
- [SYBYL] SYBYL, Version 6.5.3, HQSAR Module, Tripos Inc., St. Louis, Minnesota, USA.

- [Takahashi 1992] Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic identification of molecular similarity using reduced graph representations of chemical structures. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 639 – 643.
- [Tversky 1977] Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, 84, 327 – 354.
- [van de Waterbeemd 2003] van de Waterbeemd, H.; Gifford, E. ADMET IN SILICO MODELLING: TOWARDS PREDICTION PARADISE? *Nature Rev. Drug Disc.* **2003**, 2, 192 – 204
- [Visco Jr. 2002] Visco Jr., D.P.; Pophale R.S.; Rintoul M.D.; Faulon, J.L. Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *J. Mol. Graph. Model.* **2002**, 20, 429-438.
- [Walker 1991] Walker, P.D.; Artcea, G.A.; Mezey, P.G. Complete Shape Characterization for molecular charge density representation by gauss type functions. *J. Comput. Chem.* **1991**, 12, 220 – 230.
- [Walters 1998] Walters, W.P. Virtual Screening – an Overview. *Drug Discov. Today.* **1998**, 3, 160-178.
- [Warmuth 2003] Warmuth, M.K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 667 – 673.
- [Wilkins 1980] Wilkins, C.L.; Randic M. A graph theoretical approach to structure-property and structure-activity correlations. *Theor. Chimica Acta* **1980**, 58, 45 – 68.
- [Willett 1986] Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarity Measures of Intermolecular Structural Similarity. *Quant. Struct.-Act.Relat.* **1986**, 5, 18-25.
- [Willett 1992] Willett, P. A Review of 3-Dimensional Chemical Structure Retrieval Systems. *J. Chemometrics* **1992**, 6, 289 – 305.
- [Willett 1995] Willett, P. Searching for pharmacophoric patterns in databses of three-dimensional chemical structures. *J. Mol. Recogn.* **1995**, 8, 290 – 303.
- [Willett 1998] Willett, P.; Barnard, J.M.; Downs, G.M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 983 – 996.
- [Willett 2000] Willett, P. Similarity and Diversity in Chemical Libraries. *Curr. Opin. Biotech.* **11**, 85 – 88.
- [Xing 2002] Xing, L.; Glen, R.C. Novel methods for the prediction of logP, pKa and logD. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 796-805.
- [Xing 2003] Xing, L.; Glen, R.C; Clark, R.D. Predicting pKa by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 870 – 879.
- [Xue 1999] Ling, X.; Godden, J.W.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 881 – 886.
- [Xue 2000] Xue, L.; Bajorath J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 801 – 809.
- [Xue 2001] Xue, L.; Stahura, F.L.; Godden, J.W.; Bajorath, J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 394 – 401.