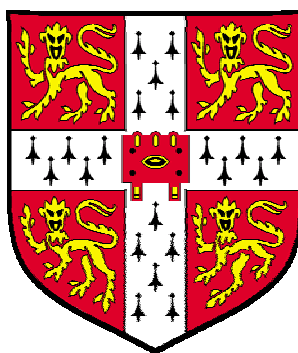


STUDIES ON MOLECULAR SIMILARITY

Andreas Bender, Darwin College

A thesis submitted for the degree of Doctor of Philosophy of the
University of Cambridge on 11 November 2005



Supervisor Robert C Glen

Ex parte enim cognoscimus.

Vulgata, Epistula ad Corinthios 1. 13,9.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

The dissertation does not exceed the word limit for the respective Degree Committee.

1 Abstract: Andreas Bender, ‘Studies on Molecular Similarity’

The “Molecular Similarity Principle” states that structurally similar molecules tend to have similar properties – physicochemical as well as biological ones – more often than structurally dissimilar molecules. The question then arises of how to define ‘structural similarity’ algorithmically and how to validate its usefulness.

In this thesis, two conceptually different approaches are implemented and explored. The first of the approaches is based on the molecular connectivity table, combining circular fingerprints (similar to ‘Augmented Atoms’), information-gain based feature selection and the Naïve Bayesian Classifier for model generation. On two standard datasets, this method outperformed many other established similarity searching methods. While both similar descriptors and Bayesian-like fragment weighting schemes have been explored previously, both the evaluation on standard datasets and the emphasis of the importance of feature selection in combination with the Bayes Classifier are novel to the work presented here.

The second of the approaches presented is based on force-field (GRID) derived energetic information about the surface of the molecule, locally capturing interaction profiles of possible ligand-receptor interactions, such as hydrophobic, charge and hydrogen bond interactions. This descriptor uses points relative to the molecular coordinates, thus it is translationally and rotationally invariant. Due to the local nature of the descriptor, conformational variations are seen to cause only minor changes in the descriptor. Compared to other approaches, it performs well with respect to the retrieval of active structures and selected features are shown to correspond to binding patterns observed crystallographically. In addition to employing force-field probes the application of quantum-mechanically derived surface descriptors (defined *via* the COSMO continuum solvation model) was explored. The screening charges calculated here also define properties relevant to ligand-target binding such as hydrogen bond donors and acceptors, positive and negative charges and lipophilic moieties from first principles. Encoding of properties is performed by three-point pharmacophores (3PP) which were found to outperform other approaches.

In the concluding chapters, two problems in using similarity approaches are illustrated. Firstly, the problem of scaffold hopping using circular fingerprints is demonstrated on an

independent dataset, namely that of the ‘HTS Data Mining and Docking Competition’ of McMaster University. The conclusions are two-fold: on the one hand the exact fragment matching similarity searching method employed here is not capable of finding completely novel hit structures; although, as shown earlier, it is able to combine knowledge from multiple active structures to give novel combinations of features. On the other hand this study emphasizes the requirement for a comparable distribution of chemical features of the training and of the test set.

Secondly, the information content of fingerprint descriptors is compared to very simple descriptors (‘atom counts’) and, depending on the dataset, a variation of none up to rather small (but significant) improvement of the performance of more sophisticated descriptors is observed. The added value of many currently used virtual screening methods (calculated as enrichment factors) over simple methods needs to be put into context of their added complexity (e.g., computational expense). The observed effect is much less profound for simple descriptors such as molecular weight and only present in the case of atypical (larger) ligands. The current state of virtual screening is not as sophisticated as might be expected which is due to descriptors still not being able to capture structural properties relevant to binding. This fact can partly be explained by highly nonlinear structure-activity relationships, which represent a severe limitation of the “similar property principle” in the context of bioactivity.

2 Acknowledgements

Firstly I thank Bobby Glen, who gave me academic guidance, freedom as well as all support I could wish for to finish this project. The initial steps of this work are based on postdoctoral work of Stephan Reiling and I would like to thank Stephan for preparing the ground for this thesis. For great help and support from the machine learning side I thank Hamse Y. Mussa.

Jonathan Goodman, John B. O. Mitchell and Peter Murray-Rust are thanked for many interesting and stimulating discussion and Charlotte Bolton for keeping the computers up and running. Susan Begg made the Unilever Centre a very friendly place for the last years and I am grateful for all help from her side.

Gavin Harper, Jérôme Hêrt, Uta Lessel and Yvonne C. Martin are thanked for providing us with datasets and for many interesting and stimulating discussions.

A number of participants at the 3rd Joint Sheffield Conference on Chemoinformatics and the 7th International Conference on Chemical Structures are thanked for input on the methods presented.

Jonathan Goodman and Peter Willett are thanked for many helpful comments on a previous version of the thesis which led to the inclusion of significant information.

For financial support I thank The Gates Cambridge Trust, Unilever and Tripos Inc.

3 Correspondence to previous publications

The majority of this thesis has been published previously in several publications in refereed journals. Please find the articles corresponding to the chapters of this thesis below.

Introduction (Chapter 6)

Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, 2, 3204-3218.

Fragment-based similarity searching (Chapter 7)

Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 170-178.

Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1708-1718.

Surface-fingerprint based similarity searching (Chapters 8 and 9)

Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT). *IEEE Int. Conf. Syst. Man Cybern.* **2004**, 5, 4553 – 4558.

Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, 47, 6569-6583.

Bender, A.; Klamt, A.; Wichmann, K.; Thormann, M.; Glen, R.C. Molecular similarity searching using COSMO screening charges (COSMO/3PP). *Lect. Notes Comput. Sci.* **2005**, 3695, 175 – 185.

Problems with molecular similarity (Chapters 10 and 11)

Bender, A.; Mussa, H.Y.; Glen, R.C. Screening for DHFR inhibitors using MOLPRINT 2D, a fast fragment-based method employing the Naïve Bayesian Classifier: Limitations of the descriptor and the importance of balanced chemistry in training and test sets. *J. Biomol. Scr.* **2005**, 10, 658 - 666.

Bender, A.; Glen, R.C. A Discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, 45, 1369 – 1375.

4 Table of Contents

1	ABSTRACT: ANDREAS BENDER, ‘STUDIES ON MOLECULAR SIMILARITY’	4
2	ACKNOWLEDGEMENTS	6
3	CORRESPONDENCE TO PREVIOUS PUBLICATIONS.....	7
4	TABLE OF CONTENTS	8
5	PROLOGUE.....	10
6	INTRODUCTION TO MOLECULAR SIMILARITY.....	12
6.1	DESCRIPTORS FOR MOLECULAR SIMILARITY CALCULATIONS	17
6.2	SIMILARITY COEFFICIENTS	30
6.3	PROPERTIES OF FINGERPRINTS AND SIMILARITY COEFFICIENTS AND THE EFFECT OF DATA FUSION.....	33
6.4	CONSENSUS SCORING	36
6.5	THE RECEPTOR IS KING.....	37
6.6	OMITTING IMPORTANT FEATURES: FREE ENERGY, ENTHALPY AND ENTROPIC DESCRIPTIONS FOR BINDING REQUIRE DESOLVATION TERMS	38
6.7	LOCAL SIMILARITY REQUIRES NON-LINEAR MODELS. OR: WHY LINEAR REGRESSION TECHNIQUES ARE NOT THE METHOD OF CHOICE.	42
6.8	SUBSTRUCTURAL ANALYSIS, BINARY KERNEL DISCRIMINATION AND THE NAÏVE BAYES CLASSIFIER.....	44
6.9	OTHER APPLICATIONS OF MACHINE LEARNING METHODS.....	49
6.10	CONCLUSIONS AND OUTLOOK	51
7	FRAGMENT-BASED SIMILARITY SEARCHING (MOLPRINT 2D) ...	52
7.1	MATERIAL AND METHODS.....	52
7.2	RESULTS AND DISCUSSION	59
7.3	CONCLUSIONS.....	91
8	SURFACE-FINGERPRINT BASED SIMILARITY SEARCHING	93
8.1	MATERIALS AND METHODS	95

8.2	RESULTS AND DISCUSSION	104
8.3	CONCLUSIONS.....	132
9	COSMO-DERIVED SCREENING CHARGES AND THEIR RELEVANCE FOR MOLECULAR INTERACTIONS.....	135
9.1	MATERIAL AND METHODS.....	137
9.2	RESULTS AND DISCUSSION	140
9.3	CONCLUSIONS.....	144
10	RESTRICTIONS OF THE MOLPRINT 2D CIRCULAR FINGERPRINT WITH RESPECT TO SCAFFOLD HOPPING, ILLUSTRATED ON THE ‘HTS DATA MINING AND DOCKING COMPETITION’ DATASET.....	145
10.1	MATERIAL AND METHODS.....	145
10.2	RESULTS AND DISCUSSION	146
10.3	CONCLUSIONS.....	154
11	ON THE INFORMATION CONTENT OF MOLECULAR DESCRIPTORS WITH INCREASING LEVEL OF SOPHISTICATION	156
11.1	MATERIAL AND METHODS.....	160
11.2	RESULTS AND DISCUSSION	161
11.3	CONCLUSIONS.....	169
12	EPILOGUE	171
	REFERENCES.....	172

5 Prologue

This thesis was meant to be a major step in my personal interest in molecular similarity – and in some sense it was, but in others it was the cause of some disappointment. This interest started with an internship and part-time work at a cheminformatics company, CallistoGen AG, from summer 2000 on. Guided by Li-hsing Wang (now with Sanofi-Aventis) I was introduced to this topic – and, while he did tremendous work in explaining concepts to me and was very supportive, I still did not grasp the problems associated with this concept: “Molecular Similarity”.

While the first chapters of this thesis deal with approaches to measure the similarity of molecules, part of it also deals with the shortcomings of the descriptors we use, namely chapters 10 and 11. In these chapters the information-content of very simple descriptors (‘atom counts’) is compared to that of more sophisticated fingerprints, and we do not see in all cases, that the more sophisticated descriptors really contain more information relevant to the problem (chapter 11). Also (but this is more a descriptor-dependent definition) the problem of scaffold hopping is demonstrated, namely on the HTS Data Mining and Docking Competition dataset (chapter 10).

In approaching the project I underestimated shortcomings and underlying assumptions of the ‘molecular similarity principle’. Some of them are that

1. Similarity of molecules is both context (e.g. receptor) and location-dependent.
2. Current descriptors treat molecules as static entities – but even by definition receptor binding involves dynamical motions of the protein.
3. No hypothesis exists which kind of interaction of the ligand and the receptor actually causes (for example agonistic or antagonistic) action: Is it occupancy? Is it on-off rates? Or some completely different property? This reflects back on the question which properties of molecules cause similarity to the system (i.e. the same biological response).
4. Similarity (in the context of bioactivity) is clearly a non-linear problem, as illustrated by the recent success of for example k-NN QSAR¹. Descriptors don’t capture this, and learning algorithms often fail to model nonlinearity appropriately due to the paucity of data (relative to the daunting size of chemically accessible space²).
5. Multiple binding modes and / or even multiple binding sites often invalidate the concept of ‘similar molecules = similar effect’.

6. Protein-ligand binding is the result of the difference of two large numbers and thus a delicate equilibrium position. Simple treatment such as ‘there is a hydrogen bond donor in molecule A and one in molecule B and therefore they are similar’ neglects subtle differences in solvation and desolvation, rendering one interaction favourable while the other is not.
7. Cooperativity in binding³ is a well-known event – current modelling tools, be it similarity- or docking-based ones, fail to appropriately model this observation appropriately.

Many of these questions are discussed in the following introduction. Here they should only illustrate that there is a lot we do not understand about biomolecular systems and that perfect classification in a high-dimensional and non-linear chemical space with the light switched off cannot be achieved throughout with the methods currently at our disposal.

Finally I would like to thank my mentors (in chronological order) Li-hsing Wang, Gisbert Schneider and Bobby Glen for their academic and personal support throughout the years. I wouldn’t have been able to achieve what I did without them.

6 Introduction to molecular similarity

Many “rational” drug design efforts are based on a principle which states that structurally similar compounds are more likely to exhibit similar properties⁴⁻⁸. Indeed, the observation that common substructural fragments lead to similar biological activities can be quantified from database analysis^{9,10}. By extension from the molecular graph to molecular properties, this leads to a concept; molecular similarity, which is a term widely used in the chemical literature^{5,11}. Similarity methods have found particular favour in the pharmaceutical industry¹²⁻¹⁴. Indeed, medicinal chemistry relies heavily on the concept of bioisosterism in which similar substructures may be interchanged whilst maintaining some degree of activity^{9,15}.

Reasons for the increasing popularity of similarity based methods include technological advances in high throughput screening and synthesis which have taken place over the last decade and resulted in the necessary application of computer based methods for compound selection and evaluation to a much greater degree than before. In tandem, computer power has dramatically increased, enabling similarity applications to be performed on very large databases of molecules. Driving the introduction of these new applications is the desire to find patentable, more suitable lead compounds as well as reducing the high failure rates of compounds in the drug discovery and development pipeline¹⁶. Fast, early and reliable prediction of suitable/unsuitable candidate structures is crucial.

Nonetheless, there are also natural limits of molecular similarity methods. As soon as the amount of information one possesses about one particular problem increases (e.g. about a receptor), the advantage of molecular similarity methods (that no external knowledge is necessary) eventually becomes a limiting factor, since taking advantage of this additional knowledge may suggest an alternative approach such as e.g. ligand protein docking. As a general rule, the molecular similarity concept is most often applied when knowledge of the system is sparse.

Also, as a result of negative public opinion with respect to animal testing, *in-silico* methods are seen as one way to reduce *in-vivo* testing. An additional driver is legislation, particularly in Europe where home and personal care products in the European Union will not, starting from 2009¹⁷, be tested on animals. Companies will then have to rely to a greater extent on their in-house compound libraries already tested for safety and may apply molecular similarity methods⁴ in order to find compounds possessing the desired

safety profiles. This of course raises the spectre of untested toxic synergistic effects occurring in novel formulations of known compounds. Computer modelling of such pharmacodynamic effects is at a very early stage and molecular similarity will probably have a role to play in evaluating risk¹⁸.

To illustrate the steady growth of this area over the last two decades, the number of publications indexed by the Web of Knowledge¹⁹ containing “molecular similarity” in the title or in either the title, abstract or keywords is shown in Figure 1.

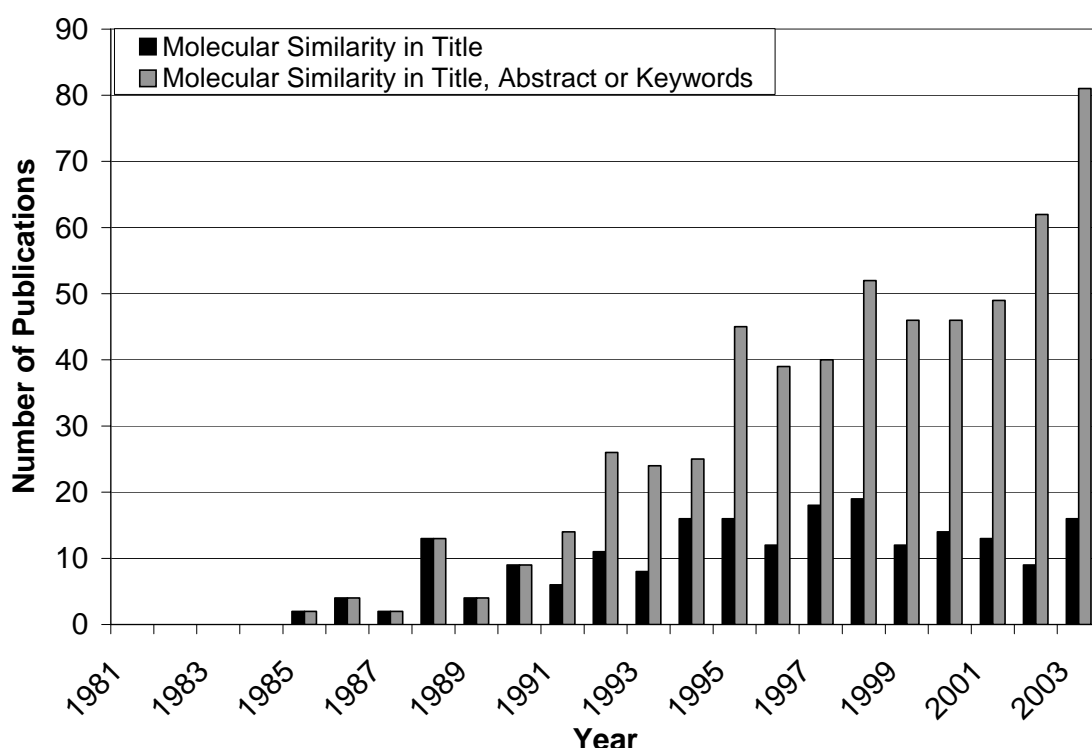


Figure 1. Number of publications indexed by the Web of Knowledge per year which contain “molecular similarity” in the title (black) or in title, abstract or keywords (grey).

Most researchers will be aware that the absolute number of publications using or citing molecular similarity methods is growing steadily. In addition it is interesting to observe that this field is maturing to become an “established” *modus operandi*. The ratio of publications containing molecular similarity in the title, abstract or keywords to the number of publications containing the words only in the title is expanding, with the ratio being 1 in the years until 1990 to about 4 in 2003, reflecting more applications of similarity based methods compared to method development.

Molecular similarity is a dynamic and evolving area of research and has been regularly reviewed. Johnson and Maggiora⁵ and Dean¹¹ wrote comprehensive books in this area.

Recently, books by Leach and Gillet²⁰ and Gasteiger²¹ have included sections on molecular similarity. General reviews of molecular similarity are given by Willett⁴, Walters *et al.*²², Gillet *et al.*²³, Bajorath²⁴ and Bender and Glen⁸. A good critique, particularly of the misuse of similarity measures is presented by Nikolova and Jaworska²⁵. A justification for the large number of molecular similarity methods is given by Sheridan and Kearsley¹². Bajorath discusses the role of similarity in the integration of *in silico* and *in vitro* screening²⁶, while Johnson *et al.*²⁷ attempts to characterize similarity methods (at least those known at that time). Some caveats of molecular similarity such as different mechanisms of action and target-dependent similarity are discussed by Kubinyi²⁸. Finally the reader is referred to Tversky²⁹, who describes early approaches to similarity in psychological testing which have been adopted by later researchers to describe similarity in molecules. Of interest here is that similarity assessments are influenced by ‘context, perspective, choice alternatives and expertise’³⁰. The choice of features, transformations and structural descriptions to describe entities (molecules in our case) will govern the predictions made by similarity models as much as do the model’s mechanisms for comparing and integrating these representations.

The fundamental observation that we can derive from these facts is that *similarity has a context*. Two vials of a yellow compound may be very similar in colour (absorption spectrum) but wildly different in biological activity. How far the context of a particular similarity argument can be taken (the ‘neighbourhood effect’) also depends on the discontinuities found in receptor-ligand interactions – clearly, the similarities studied are seldom linear and often have major discontinuities.

A brief overview of approaches to molecular similarity is given in the second part of this perspective, the emphasis being on representation of molecules in chemical space.

From Figure 1, it is obvious that there exists a great deal of information about both the methods and applications of molecular similarity. Now, as the discipline matures, it is timely to evaluate methods thoroughly. This could be approached by evaluating identical data sets using different methods. Unfortunately, different data sets or a small number (or even single) data sets are often used in the evaluation of (particularly) new algorithms, so performance is not directly comparable. In the best interests of researchers in this area, in order to make results comparable, it would be helpful to agree on some standard data sets for the prediction of different target properties. Indeed, it has been suggested that in order for a new method to be thoroughly tested, at least ten diverse sets should be used¹². One recent dataset that falls into this category was published by Hert *et al.*³¹. Additionally,

because direct comparison of performance is difficult, a different route can be explored in which the underlying assumptions of molecular similarity methods are examined. Corroborating evidence, ideally based on mechanism of action, can then be sought in the literature.

An important point is that to *accurately* define similarity on the basis of ligands alone is impossible (or futile) as the presence of external determinants or perturbations of the binding mode (and indeed e.g. diffusion rates, entropic contributions, desolvation, multiple binding modes and pharmacodynamics) of the ligand interaction with the receptor are (conveniently for many applications) unknown. In particular, the use of single valued biological measures (e.g. IC_{50} or K_i) is often a gross approximation of the ligand binding event being studied – dose response curves are hardly ever collinear with different pharmacodynamics and K_i values are often mixed with different displaced ligands being used in the same dataset. We will also discuss the desolvation energy and its possible problems for ligand based similarity. Furthermore, the importance of local similarity models (island models) and the problems of linear, one-model QSAR approaches are examined.

Some approaches, such as comparative molecular field analysis³² (CoMFA) or quantum molecular similarity^{33,34} require alignment of molecules, which is difficult to perform in cases of molecules with substantially different structures. For this reason, there have recently been developed descriptors which attempt to circumvent alignment³⁵⁻³⁷. Some of these descriptors and analysis methods possess the property of back-projectability of features from descriptor space to geometrical space. Thus, it is possible to generate human-understandable models (such as “a hydrogen acceptor 7 Ångstroms apart from a hydrogen donor is crucial”) and also to optimise structures according to the model.

Generating descriptors has a long history³⁸⁻⁴⁰ for molecules and is often the first step in molecular similarity methods. A distance measure of molecular representations in ‘chemical’ space is a required second step. This is often performed using association, correlation or distance coefficients which are based on the presence/absence of binary representations of features in the molecule (e.g. molecular substructures). The Jaccard Coefficient⁴¹ (also known as Tanimoto Coefficient) is the most widely used in practice. This coefficient of similarity was originally introduced in 1908 to measure the similarity of populations of biosystems and was later applied to the calculation of molecular similarity. Several dozen similarity coefficients are known^{4,42,43}. Over the last years it has emerged that binary representations of molecules in combination with similarity

coefficients possess some implicit properties which skew the results of similarity searches and may introduce unintentional weighting with respect to e.g. size^{44,45}. The user of those algorithms has to be aware of implicit and underlying tendencies of binary bitstrings as well as similarity coefficients, and this is discussed later. In particular, these methods will often find similarity in common substructural features which may not be reflected in how the receptor actually recognises the molecules⁴⁶. Small changes in structure can result in small changes in fingerprint similarity but large changes in molecular properties – or how a receptor perceives the ligand. Patterson *et al.* discuss this in great detail in their well-known paper on neighbourhood behaviour⁶ the essence of which is also illustrated in Figure 2. More recently, an extension of this principle in the context of multiple ligand-target interaction profiles has been presented⁴⁷.

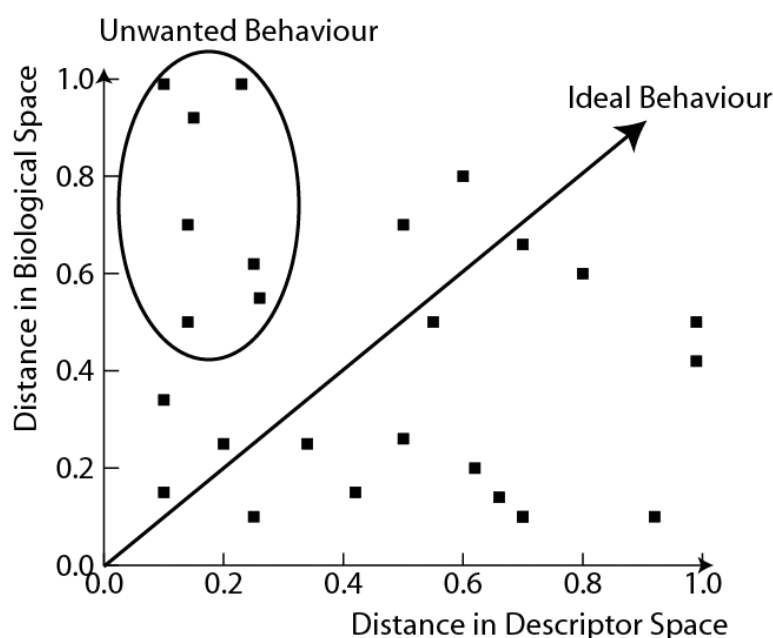


Figure 2. Illustration of neighbourhood behaviour which defines the relationship between distance in biological space (e.g. activity difference) and difference in the distance in descriptor space. If a descriptor places compounds in the upper left hand area of the graph it means that compounds show little change in descriptor space but large changes in biological space. This behaviour is even less desired than compounds showing large changes in descriptor space but small changes in activity space which may be a result of e.g. different binding modes. Ideally, distance in biological space and distance in descriptor space show a linear relationship.

Indeed, the question of how similar molecules have to be to display neighbourhood behaviour using a particular descriptor^{6,7} (to be useful in similarity searching) is dependant on the property (biological activity) in question, and in particular on the linear behaviour of the relationship between descriptor and biological activity. The studies of

Martin *et al.*^{7,46} demonstrate that compound libraries which commonly contain series of analogs which are designed to act at a particular receptor (as is found in drug discovery programs) will indeed have a higher proportion of active neighbours – obviously they will often find their own analogous series or molecules synthesised specifically to act at that receptor. This skews performance evaluations of virtual screening experiments in favour of 2D descriptors. Their study suggests that the real efficiency of neighbourhood based virtual screening (at least using Daylight fingerprints with a Tanimoto measure of similarity) is nearer 30% and indeed 5-10 similar compounds to each probe molecule need to be tested to have a high probability of finding actives. The screening results are better than random selection; however, this is a particularly important study if we are not to fool ourselves into believing better numbers due to unfortunate experimental design. (This also relates to the concluding chapters of this thesis.) Also, it is clear that the performance measurement of these methods is critically dependant on the testing methods and molecules used in the dataset – and also of the questions that are being asked. It is easy to skew these results to give a favourable outcome. Similarity is, after all, a ‘fuzzy’ concept⁴⁸.

We shall give an overview of molecular descriptors used in similarity calculations in section 6.1, followed by a short summary of similarity coefficients currently used (section 6.2). Section 6.3 deals with implicit properties of binary representations of molecules while the following section 6.4 describes the concept of ‘consensus scoring’. A criticism of one of the important aspects of similarity calculations is given in section 6.5: that similarity possesses a context; it is determined by the environment within which it is perceived (or computed). Chapters 6.6 and 6.7 discuss two possible pitfalls with similarity calculations: that desolvation effects are often not additive in a simple fashion and are generally neglected (thereby overemphasizing electrostatic complementarity) and that linear methods are not capable of modelling a relationship between descriptors and activity if the underlying activity space is rugged. Applications of machine learning methods in similarity calculations are given in chapter 6.8, before we conclude and give an outlook in chapter 6.9.

6.1 Descriptors for molecular similarity calculations

A very large number of descriptors have been developed that can be used in similarity calculations. They are typically designed to provide a molecular description that is transferable, in an information-preserving representation, to an abstract descriptor space.

Molecules are compared in three steps: representation of molecular structures in chemical space, feature selection (this step is optional) and comparison of structures. This section deals with recent developments of the first step, molecular representation in chemical space, which is also known as the generation of molecular descriptors. A general overview of molecular descriptors is given by Todeschini and Consonni⁴⁰.

A descriptor places two molecules in chemical space at a distance that reflects their similarity in this particular descriptor space. The ideal descriptor places two molecules at a distance that is proportional to their differences with respect to bioactivity, physicochemical or any other of their properties of interest. As two molecules may be perfectly similar with respect to one property (e.g. molecular weight) but completely different with respect to a second property (e.g. lipophilicity), it becomes clear that there cannot be one similarity measure and one descriptor that correlates with every molecular property at the same time. In different “similarities”, different features emerge as being important (and in our case, different bioactivities invariably require different descriptors). One-dimensional property descriptions assign only one number to the molecule. A number is usually derived from computed physicochemical properties (e.g. polarisability, volume, molecular weight). Since no geometrical information is contained in the descriptor (e.g. in which conformation a molecule would interact with a receptor), they are often employed for the prediction of physical properties or as a more general property (such as miscibility) that can be associated with properties required for receptor binding. Examples using such descriptors are clustering of compound databases⁴⁹ and database comparisons⁵⁰⁻⁵³. It should be noted that, in practice, one commonly assigns a large number of one-dimensional descriptors for similarity calculations because the relevant property is in most cases unknown and can only be – that is hoped - empirically approximated as a combination of different one-dimensional properties.

One-dimensional linear representations attempt to represent the molecule as a linear string where nodes represent atoms (or groups of atoms). This can be compared to the representation of proteins using one-dimensional sequences of amino acids. To compare molecules, algorithms similar to protein sequence alignment can be applied to compare two molecules⁵⁴. This one-dimensional representation of molecules as one-dimensional chains was recently extended to be able to utilize multiple query molecules⁵⁵.

Topological indices and other graph-based descriptors constitute the next group of descriptors. Topological indices are integer or real-valued numbers that are derived from the connectivity matrix and may contain additional property information about 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^{56,57}. As yet, there are hardly any applications published using third-generation topological indices⁵⁸.

Several hundred alternative topological descriptors have been published⁵⁹⁻⁶¹. One important aspect of topological indices is that they are derived solely from the connectivity matrix of a molecule and thus do not consider conformational variability and three-dimensional structure. For a recent review on topological indices, see Balaban⁶² and Estrada and Uriarte⁵⁸. Hall and Kier⁶³ 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”⁶⁴.

Another group of descriptors are fragment or substructure based descriptors. Maximum common substructure (MCS) searches are among the earliest substructure searching algorithms used (see e.g. Cone *et al.*⁶⁵, although the concept was already used much earlier). They are often employed to find substructures ‘similar’ to a template substructure e.g. to find all those structures containing an ethylamine fragment. This would find e.g. piperazines and piperidines as well as ethylamines. One has to be aware that substructure searching implies perfect matching of molecular connectivities and atom types (instead of using some kind of fuzzy concept to define actual similarities). These searches tend to be time-consuming due to the NP-complete (no NP-complete problem can be solved in polynomial time) nature of the problem which in the worst scenario becomes an exhaustive search. Developments in substructure searching can be found in a recent review⁶⁶. Substructural analysis (comparing the presence or absence of substructural fragments to a biological activity) in a simple form is often dubbed Free-Wilson-Analysis as Free and Wilson published one of the early applications of this type⁶⁷, followed shortly afterwards by Cramer.⁶⁸ This is still an active area of research^{23,69}. Ghose and Crippen developed counts of a selected set of 110 fragments based on carbon, hydrogen, oxygen, nitrogen, sulphur and halogens as descriptors⁷⁰ which were initially applied to the prediction of the octanol/water partition coefficient (logP). Rarey and Dixon⁷¹ represents molecules as tree-like structures called “Feature Trees” where nodes are assigned putative interaction types currently derived from FlexX interaction profiles. Feature Trees are

closely related to reduced graphs^{66,72-74} in that sets of atoms may be collapsed into single nodes, but the methods employed to compare Feature Trees and molecular graphs show a wide variety of approaches. Other fragment-based molecular descriptors are “molecular tree” fingerprints^{75,76} or related “Atom Environments”⁷⁷⁻⁸⁰ (MOLPRINT 2D) which are discussed later in this thesis. “Mini fingerprints” also contain bits which denote the presence or absence of fragments⁸¹⁻⁸³. For both “molecular tree” and MOLPRINT 2D fingerprints, fragments spanning two bonds from a central atom (five atoms in diameter) were found to be most effective in similarity searching⁷⁹ and QSAR studies⁷⁶.

Among the most popular methods for molecular similarity searching are 2D fingerprints (those based on connectivity information only) since they are fast to compute and information-rich with respect to distinguishing different bioactivity classes⁸⁴. Molecules are represented by binary (bit) strings which show the presence and absence of molecular features in a particular structure. Those features may either be based on predefined fragments or encode the presence/absence of atom pairs at a given distance along the molecular scaffold. Predefined fragments are more precisely referred to as “structural keys” while the latter approach of encoding molecular paths is closer to the original meaning of “fingerprints”. 2D representations which belong to the first group are for example MDL keys⁸⁵ which encode, in the publicly accessible key set, 166 different structural features. (Other examples were also given in the preceding paragraph.) Actual “fingerprints” are for example Daylight fingerprints⁸⁶ which encode all atom pairs up to seven bonds apart. Since the number of possible atom arrangements is virtually infinite, in the latter case no pre-defined bits can be assigned to each feature. Instead a process called “hashing” is performed, where (multiple) features are assigned to bits of the fingerprint which is typically a few thousand bits of length. In this transformation step feature “clashes” are unavoidable and equifrequency of setting bits is desirable to convey maximum information. Another popular fingerprint definition are Unity⁸⁷ fingerprints which contain different blocks of information, including pre-defined structural features, the presence of ring systems, and also a block encoding hashed information similar to the Daylight definition. Apart from the question how features are precisely mapped onto fingerprint bits also the encoding of atoms can be varied from element atom types to those attempting to represent putative ligand-target interactions such as hydrogen bonding capabilities, charges and lipophilic moieties.

Descriptors derived from combinations of other descriptors, using correlation or principal components methods are also popular. BCUT descriptors^{88,89} derived by diagonalising a

matrix of atom-based properties to generate a set of new decorrelated descriptors (based on the smallest eigenvalues) have found use particularly in compound selection where diversity in compound library design is desired.

Using spatial (3D) information, popular methods, such as CoMFA³², use data derived from a series of overlaid molecules to create a model. Often molecular alignment is performed in order to overlay substructures which are thought to have similar properties, placing corresponding interaction groups in neighbouring areas in space. Using 3D information thus has a price – in the absence of information on relative binding orientations, one has to be deduced or inferred. In this context, it is often better to overlay interaction points rather than structure – the overlay step is potentially the most difficult⁹⁰. Alignment of very dissimilar structures (different scaffolds) is not a trivial task. This is illustrated in Figure 3 and Figure 4. While alignment of two Angiotensin Converting Enzyme inhibitors of very similar size is possible without problems (Figure 3), alignment of two structures of different size – which still both show the inhibition of ACE *in vivo* – gives close-to arbitrary alignment (Figure 4).

An explicit method for molecular alignment is presented by Kotani and Higashura⁹¹. The smaller of two molecules is superimposed on the larger molecule. Then, nearest atomic distances of each atom of the smaller molecule to any atom of the larger molecule are calculated. Almost comparable results to the approaches by Meyer and Richards⁹² and Good⁹³ are obtained with respect to molecular similarity, but the new method is several orders of magnitude faster.

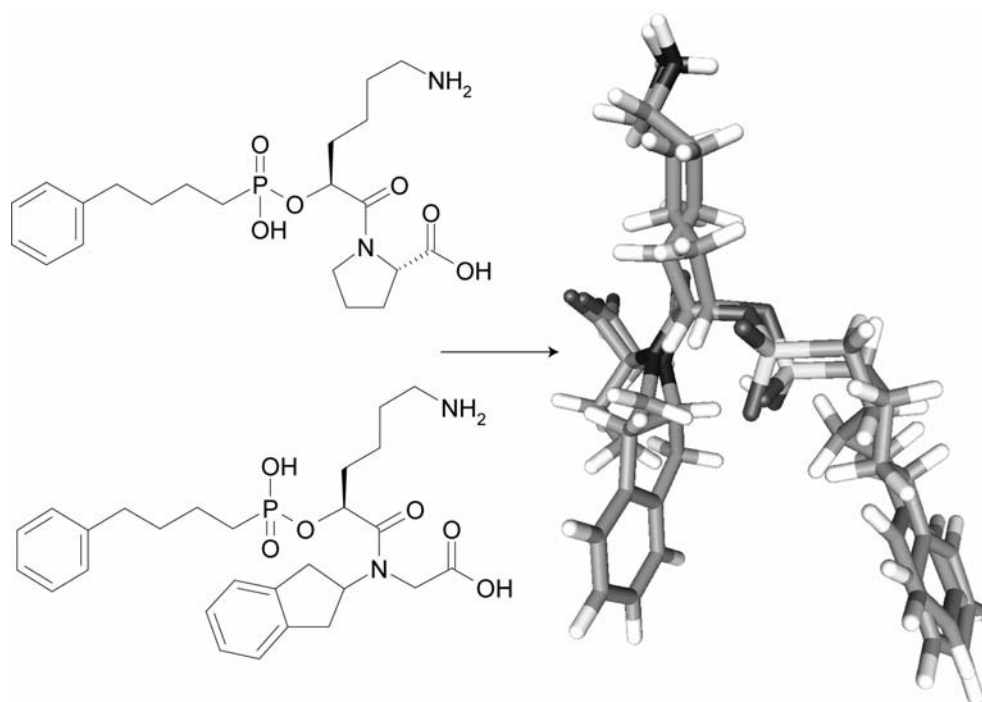


Figure 3. Alignment of two inhibitors of Angiotensin Converting Enzyme (ACE). Both molecules are of comparable size so alignment is, although still time-consuming, feasible in an unambiguous manner.

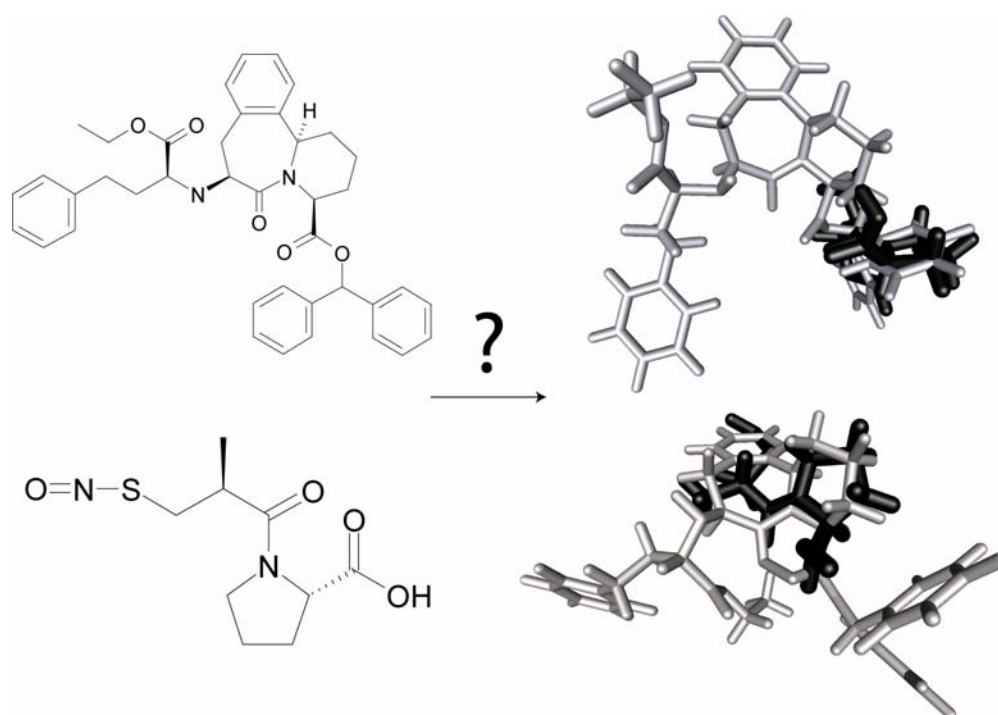


Figure 4. Alignment of two inhibitors of ACE of very different size. Alignment is in this case not feasible in an unambiguous manner. These results were obtained using a genetic algorithm (GASP) for alignment but this problem exists independent of the particular algorithm that is applied.

One novel and interesting way to remove the structural overlay criterion is to have a set of consistent ‘expert’ rules to provide a conformational and orientational overlay from which field-based analysis may be performed. This new approach, called ‘Topomers’, after their description of topological shape, was taken by Cramer and Andrew^{37,94} and appears to be surprisingly robust. Several datasets have been analysed and an additional advantage is that very large chemical libraries can be analysed quickly using a combinatorial approach to shape matching.

Quantum Similarity, which establishes the similarity of molecules based on measured derived from the wave function, such as electron density, was introduced in the early 80’s³³. Hodgkin and Richards⁹⁵ later introduced a related index that took into account not only electron distribution but also electron density. Walker *et al.*⁹⁶ and Good *et al.*^{93,97} replaced the grid approach with a Gaussian approximation leading to significant increases in performance. Furthermore, this solved problems with local minima while performing molecular alignments. The Gaussian representation has later been generalized to describe molecular shape⁹⁸. For a comprehensive review on Quantum Similarity, see Carbo-Dorca and Besalu³⁴, and for latest developments see more recent reviews⁹⁹⁻¹⁰¹. A basic introduction to the topic has been published¹⁰² and for a discussion of the significance of QM methods in similarity (particularly atoms in molecules theory) see Nikolova and Jaworska²⁵ and Boon *et al.*¹⁰¹. Quantum mechanical methods of similarity hold the promise of a better representation of structure and importantly, perturbations in molecular properties that can only be determined by evaluating the response of the wave function. Increases in computer power with speedup due to algorithm developments are making these methods more attractive in principle, although there are very few large-scale applications yet.

Many grid based descriptors based on non-quantum mechanical methods owe their inspiration to Goodford’s work in field based methods¹⁰³ in combination with a robust statistical method for variable selection (partial least squares, PLS)¹⁰⁴, were introduced in the late 1980’s with the Comparative Molecular Field Method, CoMFA³². This method was also the basis of the Comparative Molecular Similarity Analysis (CoMSIA) approach^{105,106} developed by Klebe *et al.*

Another group of descriptors makes use of the concept of the receptor and ligand as a ‘lock and key’ in which common interacting groups are found at similar distances apart. This is the pharmacophore hypothesis¹⁰⁷ and in many applications it involves identifying key functional groups and their conformationally dependant inter-fragment distance

ranges. These methods do not rely on molecular alignment; instead, relative internal distances of the molecule are used (and, indeed, one of their advantages is that no alignment is necessary). They are often referred to as multiple-point-pharmacophores. Two-point pharmacophores^{108,109} (2PP) are also known as atom pairs and represent all possible pairs of atoms in the molecule, three-point pharmacophores¹¹⁰⁻¹¹⁵ (3PP) allow for a more detailed representation of interatomic distances and four-point pharmacophores^{116,117} (4PP) are able to distinguish between stereoisomers. A potential shortcoming of 4PPs is the large number of possible combinations of interaction points and distance ranges. Where three-point pharmacophores with six interaction types and ten distance ranges give rise to 33,000 possible descriptor bits, four-point pharmacophores with the same number of distance ranges and interaction types need about 13 million bits. Four-point pharmacophores were used to examine selectivity between Thrombin, Factor Xa and Trypsin, which are all homologous serine proteases¹¹⁸. While three-point pharmacophores were not able to identify features responsible for selectivity, four-point pharmacophores appeared to identify important features. Nonetheless, the data set only comprised a very small number of molecules and further studies are necessary. Four-point pharmacophores have also been used for library design¹¹⁹ based on the Ugi condensation and a serine protease active site.

Nonetheless, this method has not yet found as wide an acceptance in the scientific community, possibly attributable to the much larger amount of information, with bit strings about 300 times as long as in the case of three-point pharmacophores (which are then about 13 Mbits or about 1.5 Mbytes long). Although four-point pharmacophores describing differences of the binding sites of different serine proteases were found¹¹⁸, it may in practice pose a problem to select a few hundred relevant features out of several million, even more so given the dependence of the method on conformational changes.

The surface-based group of descriptors focuses on the commonly accepted assumption that ligand-receptor binding is mediated by the molecular surface, e.g. by the complex shape of the van der Waals surface. Clark discusses the applicability of surface-based descriptors as a matter of principle as well as possibilities to calculate molecular surface properties¹²⁰. Some examples of the utility of surface based descriptors follow. Gaillard *et al.*¹²¹ devised a method to describe molecular lipophilicity potential and validated it by predicting logP values. Stanton and Jurs¹²² introduced the concept of “charged partial surface area structural descriptors”. The Compass method¹²³ by Jain *et al.* takes several molecules and several conformations into account and requires a user-defined interacting

pharmacophore guess. This approach has also been used for selecting library subsets in its extension called Icepick¹²⁴, where several conformations of the molecules to be compared are calculated and the three-dimensional structures are docked into each other. Jain also introduced the concept of “morphological similarity”¹²⁵ 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 required. A novel method for classifying similarity of molecules is performed by using hash keys of the molecular surface, compared to a panel of reference compounds¹²⁶. Applied to several data sets, the description was found to capture enough information for the prediction of ADME properties and target binding. Hash codes are generally used for structure storage and retrieval¹²⁷ and here, they are applied to structure-activity relationships.

Affinity-fingerprint based descriptors compare the complementarity of a ligand to a panel of reference receptors and score 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. 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 crucial for finding ligands *in vivo*, so that more meaningful results could potentially be derived from this approach. *In vitro* fingerprints were first introduced by Kauvar *et al.*¹²⁸ and shortly afterwards followed by their *in silico* counterparts^{129,130}. The latter were, for example, employed in library design¹³¹. For a recent review see Briem and Lessel¹³².

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 mapped onto descriptor space, e.g. by calculating its zero crossings. The earliest work in this area was done by Soltzberg and Wilkins¹³³, 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) approach¹³⁴. The first descriptor calculated from the vibrational spectra of molecules is the EVA descriptor¹³⁵. Here, fundamental frequencies of the vibrational spectrum are calculated and used for the comparison of molecules. A different approach¹³⁶ defines fuzzy peak areas to derive molecular features from an infrared spectrum, followed by principal components analysis.

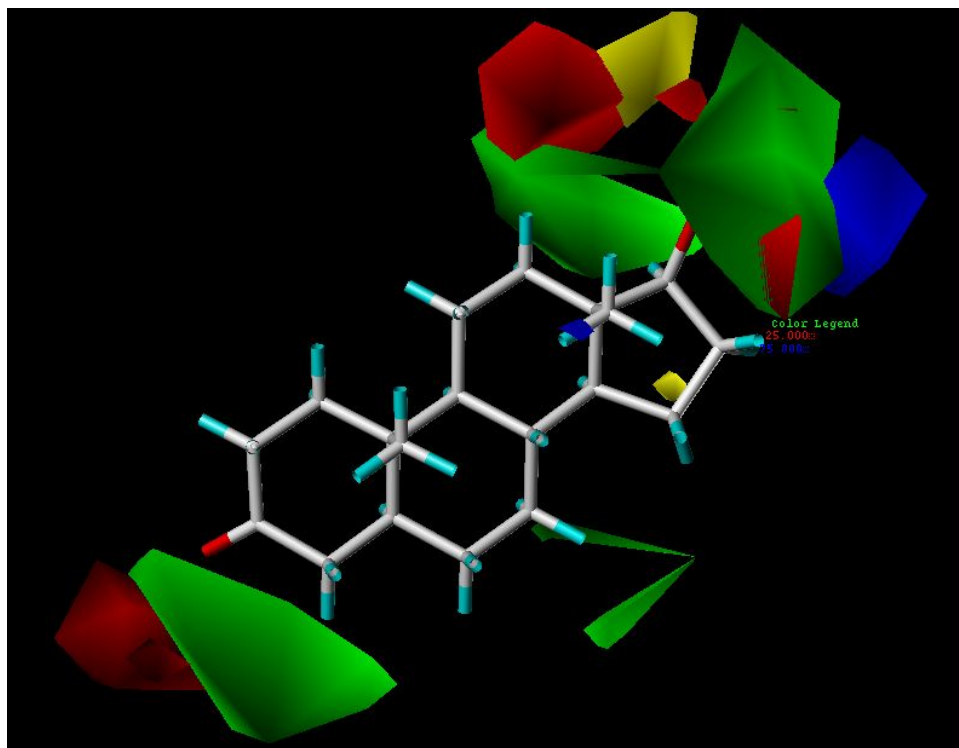


Figure 5. Areas in space projected back on dihydrotestosterone which are favourable or unfavourable for binding with respect to steric effects (green/yellow) or where negative (blue) or positive (red) charges increase activity.

One major advantage of recent descriptors is their translational and rotational invariance. In addition, some of the more recent descriptors are back-projectable, which means a feature that is found to be important (e.g. in terms of frequency of occurrence) can be projected back onto the molecules from which it was derived in some chemically meaningful way. Although most often used in structure activity relationships, can also be used in molecular similarity calculations. In the following, essentially proof-of-concept calculations on small datasets were presented by the authors and their application in the future will either prove their usefulness in practice, or not. Also, one should be aware that datasets containing rigid structures (such as the CoMFA dataset) enable alignment of structures that is rarely possible on more common (and more flexible) datasets.

An illustration of back-projectability is given in Figure 5 and Figure 6. In Figure 5 CoMFA³² was applied to 21 steroids from the well-known steroid dataset¹³⁷. The steroids were aligned and, using PLS, areas in space which contribute favourable or unfavourable to binding, either in a steric or electrostatic sense, are highlighted using colours. Thus directed structural optimization may be performed. It should be mentioned that the steroid dataset is particularly amenable to CoMFA due to the rigidity of the steroid core.

In Figure 6 features responsible for binding of a statin to 3-hydroxyl-3-methylglutaryl-Coenzyme A (HMG-CoA) reductase were back-projected on the molecular surface. Features were selected using surface point environments, information-gain feature selection and a Naïve Bayesian Classifier¹³⁸. They correspond to experimentally determined binding features¹³⁹ and correctly identify the CoA binding pocket and the lipophilic moiety. This information about binding can be used to design novel active entities.

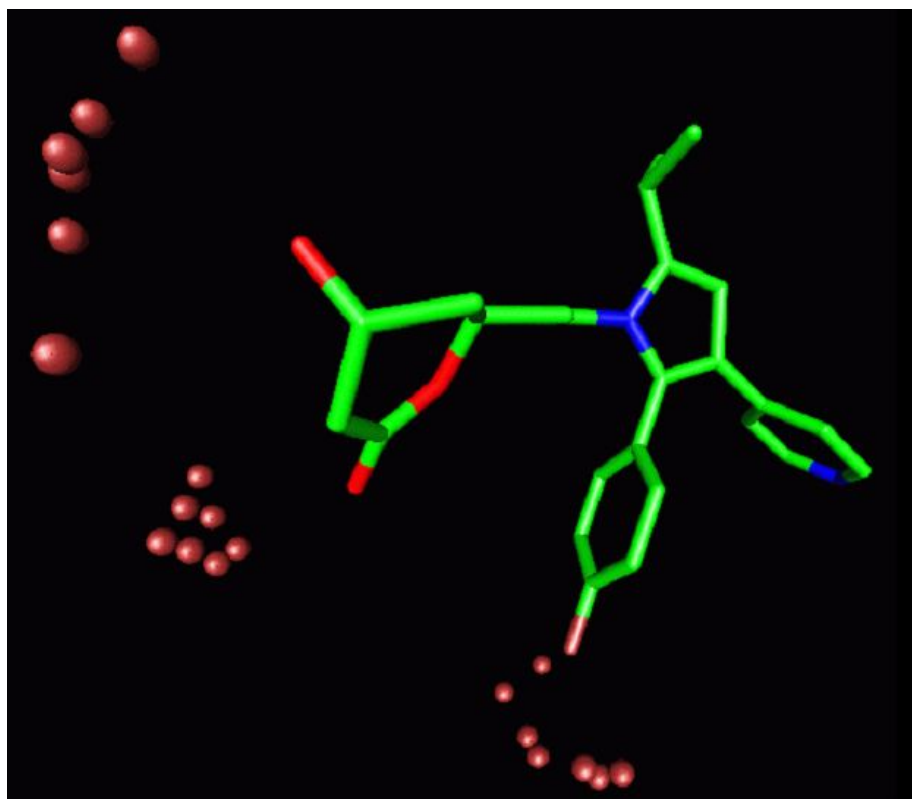


Figure 6. Features identified as being responsible for binding of a statin to HMG-CoA reductase. Employed were interaction energies of surface points which were encoded in a binary format, information-gain based feature selection and the Naïve Bayesian Classifier (MOLPRINT 3D method). Features binding to the CoA site as well as to the flexible lipophilic pocket are correctly identified. Note that while the compound shown is a prodrug (the lactone ring is ring-opened in water) the derivation of features is still correct, since in the training set a large number of prodrugs was present.

The first of the recent descriptors possessing both translational and rotational invariance and back-projectability was published by Pastor *et al.*³⁵ and called the GRIND (Grid-INDependent) descriptor. Molecular Interaction Fields (MIFs) are calculated at regularly spaced grid points using different probes. Often, DRY, N1 and O probes are employed to consider hydrophobic interactions and hydrogen bond acceptor and hydrogen bond donor fields, respectively. The fields are simplified and autocorrelation descriptors of the field

are calculated using binned distances of the fields. Only the highest products of molecular interaction energies are stored. Thus, information about less favourable sites is lost and it was found in practice that in not every case, the maximum product corresponds to a meaningful feature. At the same time, back-projectability is facilitated. Using a related program, ALMOND¹⁴⁰, descriptors can be back-projected in three-dimensional space. Using this approach, features responsible for steroid binding have been identified. Performance in a statistical sense is comparable to other methods, but with the added advantage that no alignment is necessary and that descriptors are easily interpretable. In one application, using GRID and GPCA, regions of the homologous serine proteases Thrombin, Trypsin and Factor Xa responsible for selectivity were identified¹⁴¹. The findings were in agreement with experimental data. More recently, a shape description was implemented as an additional “probe” of the GRID program¹⁴². Before the introduction of the shape description very similar descriptors were obtained for molecules with additional aliphatic chains since the resulting interaction fields were very weak (and virtually not present in the descriptor). The additional shape probe is able to capture those regions (which may cause steric repulsion and thus a huge difference in activity) and was shown to improve QSAR statistics¹⁴².

A related approach was presented by Stiefl *et al.*^{36,143,144} and named MaP (Mapping Property distributions of molecular surfaces). This algorithm first constructs equally spaced surface points on the molecular surface. Established methods such as MSMS¹⁴⁵ and others have been found not to give equally spaced surface points so an implementation of the GEPOL algorithm¹⁴⁶ was used. In a second step, a probability distribution function was calculated. Because binning was performed to give discrete values, this in effect was a distance dependent count statistics. The approach differed from the GRIND descriptor in that it uses the molecular surface instead of equally spaced grid points. Additionally, categorical variables are used which are then binned in a distance-dependent fashion. In the original method, hydrophobic, hydrophilic, hydrogen bond donor and hydrogen bond acceptor surface points were used. When applied to a steroid data set and a set of eye irritating compounds, results are broadly comparable to other methods. For a set of muscarinic compounds, a model could be developed despite high flexibility of the compounds³⁶. Again the advantage of translational and rotational invariance as well as back-projectability is seen.

Multidimensional QSAR approaches are an extension of QSAR algorithms using the three spatial dimensions for descriptor encoding such as Comparative Molecular Field Analysis (CoMFA). The fourth dimension describes conformational sampling of the molecule.

Originally, four-dimensional QSAR was introduced by Hopfinger *et al.*¹⁴⁷. Descriptors are ensemble averages of grid cell occupancies of each molecule and each conformation generated from a training set. From a high number of possible descriptors, usually only a small fraction (typically between 10 and 20) is selected by partial least squares. Then, genetic function approximation is employed to model a structure-activity relationship of the compounds. In the introductory paper, 4D QSAR performs at least as good as 3D approaches but gives additional information. Firstly, active conformations can be guessed. Secondly, important features can be part of the activity function although they remain constant in all compounds. In addition, conformational entropy can be estimated which may provide a different method to select compounds for lead selection.

In 4D QSAR, relative and absolute similarity measures exist¹¹⁷. Absolute measures use atoms; relative measures use grid cell occupancy descriptors. Absolute descriptors are not alignment dependent, whereas relative descriptors are. As an example, similarity calculations using D and L amino acids are given. The discussion gives some advantages of the 4D QSAR methods, but at least two of the advantages are also seen using other descriptors. Alignment is, as seen above, not necessary. Also, atom typing can, at least in principle, be changed to that found in other algorithms. The additional information of conformational sampling is definitely given in 4D QSAR compared to 3D QSAR, but one should keep in mind that not the amount of information but the signal/noise ratio is of importance here. Conformational sampling apart from covering conformational space also introduces noise into intramolecular distances, even more so in flexible molecules.

Several applications of 4D QSAR have been published in recent years. The construction of a virtual high throughput screen by 4D QSAR was applied to the design of glucose inhibitors of Glycogen Phosphorylase b, but no experimental testing of the constructed library is given¹⁴⁸. A 4D QSAR study of CYP450 inhibitors showed that that even if models give very similar predictions, the similarity of the models themselves could be surprisingly low¹⁴⁹. Cytochrome P450 2D6 inhibitors have already been subject to an earlier 3D/4D QSAR study¹⁵⁰. A study using Propofol analogues gave identical results concerning the interaction sites¹⁵¹. Not surprisingly, analogue compounds are predicted better than structures dissimilar to the training compounds. Flavonoids binding to the BzR site of GABA-A were examined by Hong and Hopfinger¹⁵².

Conceptually, the common receptor-independent (RI) 4D QSAR analysis has recently been extended to a receptor-dependent (RD) version¹⁴⁹. The statistical quality of the RI and RD versions of 4D QSAR were found to be about the same, but predictivity of the RD version was found to be superior. The receptor is conformationally sampled as well as the ligand. Although geometry pruning of the receptor is performed, conformational sampling is still computationally expensive. Receptor-dependent and receptor-independent 4D QSAR have also recently been applied to a set of nonpeptidic HIV protease inhibitors¹⁵³.

A modified 4D QSAR algorithm, compared to the Hopfinger version, is able to take local induced fit and hydrogen bond flip-flop into account¹⁵⁴. Allowing for a multiple representation of the receptor another dimension is added, giving a 5D QSAR method¹⁵⁵. This provides the possibility of allowing induced fit of the ligand-receptor complex. Adding another 'dimension' to the concept of QSAR, 6D QSAR is supposedly also able to cope with different binding modes of ligands¹⁵⁶.

6.2 Similarity coefficients

The comparison of pairs of molecules is most often performed *via* the calculation of similarity coefficients, which assign a single number to the similarity of the associated bit strings. Often normalization to similarity values in the range [0;1] is given. Similarity coefficients possess inherent properties, such as whether the same consideration is paid to features which are present and those which are absent. Also characteristic is the assigned similarity value with respect to interchange of the two structures to be compared: While symmetrical similarity coefficients assign the same similarity value to the comparison of molecule A to molecule B, asymmetrical similarity coefficients assign different values to each molecules.

Similarity coefficients in their binary forms usually describe associative coefficients which assign values of 1 for identical molecules and values of 0 for (in this particular representation) most dissimilar structures. Real-valued versions of similarity coefficients also often give values outside this range; for example the Tanimoto Coefficient for real-valued vectors gives similarity values in the range [-1/3;1]. Correlation coefficients calculate correlations between the descriptors vectors (which also give resulting values of 1 for identical, but those of -1 for most dissimilar structures) while distance coefficients give values in the range [0;∞], where smaller distances agree with more similar molecules (and values of 0 correspond to 'zero distance' between fingerprints, or identical

representations). Some frequently used associative similarity coefficients, correlation as well as distance coefficients are given in Table 1.

Table 1. Overview (only of the binary versions) of some frequently used (associative) similarity coefficients, correlation coefficients and distance coefficients. Standard abbreviations are a: number of bits only set in fingerprint A, b: number of bits only set in fingerprint B; c: number of bits in common between fingerprints A and B.

Associative Coefficients		
Name	Formula	Comment
Jaccard/Tanimoto Coefficient	$T_C = \frac{c}{a+b-c} = \frac{AND}{OR}$	Standard coefficient, possibly due to its appealing symmetry; nonetheless often outperformed ¹⁵⁷
Size-Modified Tanimoto Coefficient	$T_{SM} = (\frac{2-p}{3})T_C + (\frac{1+p}{3})T_{C,0}$ p: relative bit occupation T _{C,0} : Tanimoto coefficient of unoccupied bits	Little bias in diversity selection ¹⁵⁸ , little advantage over Tanimoto coefficient in similarity searching ¹⁵⁷
Dice Coefficient	$S_D = \frac{2c}{a+b}$	Binary version of Hodgkin quantum similarity index
Cosine Coefficient	$S_C = \frac{c}{\sqrt{ab}}$	Good performance in similarity searching ⁴³ ; Binary version of Carbo index
Correlation Coefficient		
Pearson Coefficient	$P = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	
Distance Coefficients		
Euclidean Distance	$S_E = \sqrt{a+b-2c}$	Monotonic with Hamming Distance; Bias towards small molecules in similarity search; Performs well in similarity search ¹⁵⁷
Hamming Distance	$S_H = a+b-2c$	See Euclidean Distance
Soergel Distance	$S_S = \frac{a+b-2c}{a+b-c}$	Complement of the Tanimoto Coefficient

6.3 Properties of fingerprints and similarity coefficients and the effect of data fusion

Binary bit-strings, computed from the presence or absence of molecular features are commonly compared using a similarity coefficient as a measure of similarity between structures. This is a particularly efficient method in the case of two-dimensional or other easy to calculate descriptors. In addition, binary representations are suited to computer processing so they are usually very fast.

In this context there are two interesting points which deserve attention.

Firstly, binary representations are not unbiased representations of molecules. They possess inherent properties which in effect skew results in similarity searching. For example, the presence/absence structure of bits, which in most cases do not consider the frequency of detected fragments, influences results.

Secondly, the selection of the similarity coefficient used (of which a large number exist⁴) determines the numerical value of the similarity. The question arises which coefficients perform best? On the one hand some of them are not as different as their names suggest, also, others may cluster a set of molecules into different categories. In addition, it has been suggested that combinations of predictions using different similarity coefficients may improve classification results.

The direct comparison of similarity coefficients is not subject of this article. For a comparison of the performance of different similarity coefficients, see Whittle *et al.*¹⁵⁷ (in the context of natural product databases), Salim *et al.*¹⁵⁹ (also covering data fusion) and Holliday *et al.*⁴³ (covering 22 similarity coefficients in combination with 2D fragment-based bit strings).

Bias of binary Representations

Binary representations of molecules generally represent the presence and absence of features by setting and un-setting bits in the fingerprint. The frequency of features (e.g. substructures) is not usually encoded in the fingerprint. In order to reduce size, fingerprints may often be folded (e.g. Daylight fingerprints). All these steps destroy information about the molecule, but on the other hand make fingerprints denser and smaller. Alternatively (or as an addition), pre-defined keys representing substructures of the molecules can be generated and included in the string. The adoption of Markush type entries in bit string representations, such as for the definition of hydrogen bond donors etc., allows for fuzzy matching of structural queries when looking for similar molecules.

Combinatorial effects

Binary representations of molecules already possess inherent properties, simply due to their presence/absence structure of common features. Computing the similarity of molecules using similarity coefficients usually results in a ratio of small numbers. Common fingerprints such as ISIS MOLSKEYS use 166 predefined keys and most fingerprints are not larger than 1024 bits. Thus, some ratios of small numbers are more likely to occur than others. Using fingerprints of up to 67 bits, Godden *et al.*¹⁶⁰ have shown that some ratios and thus similarity coefficients occur much more frequently than others. Using the Tanimoto coefficient, 1/3 was most likely to occur. This is also the expectation value of infinitely long bit strings where half of the bits are set. Values such as 0.25, 0.5 and 0.4 were also much more likely than other values. Random similarity values larger than 0.7 or 0.8 (values which are commonly said to define “similar” molecules”) virtually do not appear. Depending on the database, similarity coefficient distributions vary. Note that a similarity threshold of 0.8 was (depending on the descriptor used) commonly associated with a ‘high likelihood of the two compounds showing similar properties’, although the validity of this assumption has been thoroughly challenged recently⁷.

Size dependence of similarity coefficients

The size dependence¹⁵⁸ of the commonly employed Tanimoto coefficient⁴¹ (Tc) has already been known for some time, favouring larger molecules in similarity and smaller molecules in diversity selections¹⁵⁸. The reason for the size-dependence is that the size of a molecule influences the maximum Tc that can be computed for that molecule. This maximum is found if all its bits are matched by bits from the query. There are still query bits which smaller molecules cannot match, and where larger molecules show bits not set in the query, thus decreasing the Tanimoto coefficient. For larger molecules, mismatch of bits results in the same absolute change of the denominator, but due to larger absolute numbers the effect is a smaller relative change. Also, for larger molecules simply more possibilities exist to match features of a second molecule since more bits are present.

Flower⁴⁴ notes that larger molecules generate a different distribution of similarity scores when scoring a molecule database, compared to smaller queries. When larger queries are used, the Tanimoto coefficients tend to become on average larger and also possess larger variance. Thus it should be kept in mind that Tc values are not directly comparable among different queries and databases. The size dependence was also examined by Dixon and Koehler⁴⁵, who used the Tanimoto coefficient, XOR and the Euclidean Distance in

combination with ISIS Molskeys and Daylight fingerprints, applied to the RBI and Current Medicinal Chemistry (CMC) databases. The performance measure was biological target coverage. The diversity measure 1-Tc was found to have the most profound bias with respect to size. Still, this is not necessarily a problem, provided there are small compounds that are active in the database. The XOR classifier in turn slightly prefers large compounds in diversity selection. Holliday *et al.*¹⁵⁸ confirms that the Tanimoto coefficient is a poor tool for diversity selection if the Tanimoto coefficient is low, as it performs only as well as random selection of compounds. 14 similarity coefficients were assessed, determining upper and lower bounds and other characteristic properties. Self-similarity plots of libraries are used to assess molecular diversity of libraries. Again it is found that the Tc distribution depends on relative bit densities of compounds.

Attempting to rid the Tanimoto coefficient of its size bias, a size-modified Tanimoto coefficient was recently introduced¹⁶¹. This analysis suggests that using this methodology, size-dependence should be removed. However there exists no comprehensive study of its performance. Interestingly (or confusingly), in first reports¹⁵⁷ its performance was found to be similar to the original Tanimoto coefficient.

Clustering of similarity coefficients

Several dozen similarity coefficients are known in the literature. Some studies were undertaken to cluster similarity coefficients, presenting the opportunity to select measures from different clusters with the objective of improving results.

Experimentally, Unity fingerprints were used in combination with 22 similarity coefficients and the NCI AIDS and IDAlert database for similarity coefficient clustering and consensus scoring⁴³. Using an early stopping criterion, 11 clusters of indices were obtained, using a later criterion similarity indices were classified into three clusters. This research was extended¹⁵⁹ to include the MDDR database and BCI and Daylight fingerprints, giving rise to 13 clusters of similarity indices. It was also found that different coefficients perform better in certain ranges of molecular size (or bit density). The Russell-Rao coefficient was found to perform better in the case of large queries, while the Forbes coefficient performed better on small queries. The Tanimoto coefficient was outperformed in many cases, but not consistently. The good performance of Russell-Rao and the weak performance of Forbes were also observed in an application using the Dictionary of Natural Products Database (DNP)¹⁵⁷, where the Tanimoto coefficient was often outperformed by a factor of 2.

6.4 Consensus scoring

The theoretical basis of improving classification results by consensus scoring was presented by Wang and Wang¹⁶². In a hypothetical experiment, a library containing 5000 compounds was created and a Gaussian error distribution of affinities (but no systematic errors) was added. Assumptions such as independence of scoring functions were found realistic enough. Importantly, it was found that consensus scoring performs better due to the simple statistical reason that the mean value of repeated samplings tends to be closer to the true value. Performance measure was the number of misranks (pairs of structures not in the correct order), and was shown to increase continuously with the number of methods added. The authors conclude that three or four methods are best and recommend data fusion by rank or score.

A criticism of the work by Wang was given by Verdonk *et al.*¹⁶³, who empirically analysed the quality of the Goldscore, Chemscore and Drugscore scoring function for identifying active compounds and combinations thereof (rank-by-rank, rank-by-number and rank-by-vote). While, as predicted by Wang, the quality of consensus rankings also decreased in the order rank-by-number (called score in Wang's analysis) > rank-by-rank > rank-by-vote, Verdonk *et al.* stressed that, in practice, scoring functions are hardly of very similar quality for all targets, an assumption used by Wang. The empirical finding that consensus scoring hardly performs better than the best individual scoring function can thus be explained by the fact that inferior scoring functions only introduce noise into the evaluation. Nonetheless, consensus scoring generally was found to give more robust results than single scoring functions¹⁶³. This finding was confirmed when the clustering of similarity coefficients was used to select sets of similarity coefficients for consensus scoring^{43,159}. Out of seven targets in all except one (β -lactamase) improvement in classification was observed in most cases when 3 or more similarity coefficients were used. More recently the conditions in which consensus scoring improves results have also been shown on a theoretical basis¹⁶⁴ and, as demonstrated already before, performance of the individual methods as well as independence of errors are the important determinants for performance improvement of consensus algorithms.

There are different possibilities to combine information from several scoring algorithms to provide a single prediction. Using several data sets, Ginn *et al.*¹⁶⁵ used MIN, MAX and SUM rules, defining the combined prediction by the lowest, highest and average prediction of the individual methods. Here it was found that consensus scoring performed

better (and significant at $p < 0.05$) in 28 out of 30 runs, if the SUM rule is used. In practice, one can use this information to determine how to deal with multiple active structures which are known, and in the latter publication it is suggested that adding up individual scores of each pair of query and library compound improves results.

6.5 The receptor is king

Molecular similarity is a concept often used to estimate properties in biological systems or activity at receptors. A wide range of properties can be predicted such as physicochemical properties, which by their nature are often in a homogenous medium^{77,166} or NMR^{167,168} or IR¹⁶⁹ spectra. However, of particular importance for the pharmaceutical industry are properties like absorption, distribution, metabolism, excretion and toxicity (ADME/Tox)^{170,171} and bioactivity^{13,79,132,172}. These depend to a great extent on perturbations introduced by specific interactions between ligand and receptor (or transporter molecule, as for example in the case of active transport in absorption). Many of the ligand-induced perturbations (as well as more often than not the target structures themselves) are unknown and the interactions can vary in a non-linear fashion between ligands. For example, while in a certain range a more polar oxygen-hydrogen bond may increase affinity to the target due to more stable hydrogen bonding, a change to another donor/acceptor pair may favour solvation in the medium and therefore weaken the interaction.

The similarity problem itself determines the performance of the representation of the structure in chemical space. A “sensible” descriptor places two molecules apart from each other in descriptor space at a distance related to the differences in their activities or physicochemical properties. For different applications, different features of the molecules turn out to be important. Thus descriptors and the measure or calculation of similarity have to be defined on a case by case basis. Because of the discontinuous nature of ligand/protein interactions, similarity is a local effect with a ‘distance range’ within which it applies.

To illustrate the importance of an external reference, we may consider bioisosteric replacement of functional groups. If an ether linker (-O-) is replaced by an amine group (-NH-), broadly the same lipophilicity is retained. But if this group is involved in hydrogen bonding, depending on whether donor or acceptor properties are present in the receptor, several orders of magnitude of binding can be gained or lost by this replacement¹⁷³. Depending on whether the external reference referred to is lipophilicity or hydrogen

bonding capabilities, similarity or bioisostericity should therefore be computed differently.

Also, the magnitude or range of similarity always depends on the nature of the problem. For example, steroids exhibit totally different effects on the human body, where they can act as male sex hormones, female sex hormones, anabolics etc.^{173,174}. This different behaviour is mediated by very small changes to the core steroid structure and their complementarity to different nuclear hormone receptors. Thus, as overall similarity of steroids is very high, only small, local dissimilarities are responsible for different activities. In situations like this, an understanding of the mechanism of interaction and its influence on activity would be a better approach than a global similarity measure.

To conclude, molecular similarity is always a property that depends on an external criterion which defines similarity (a receptor, a physicochemical property). Similarity is not a property of the molecule itself, molecules are perceived as being similar (or different) by external ‘judges’ that are guided by natural laws (receptors, detectors of physicochemical properties). The magnitude of similarity that is required for similar activity also depends on the external reference, as illustrated by steroids, which are overall very similar but nonetheless possess very different properties¹⁷⁴. There are no absolute measures of molecular similarity and each case requires the selection of appropriate properties and classification methods, thereby also rendering all “absolute” similarity values obtained from calculations to be of relative value, depending on the particular parameters under which they were obtained.

6.6 Omitting important features: free energy, enthalpy and entropic descriptions for binding require desolvation terms

Many of the most interesting similarity comparisons are between molecules that could potentially bind to a biological receptor. If a ligand finds its way to the appropriate target, it is not just the non-covalent interaction of two molecules which become one more or less stably bound aggregate that determines biological effects. Before recognition and binding of the ligand occurs, both binding site and ligand have to be stripped of their solvent shells, some water molecules may remain and form bridging hydrogen bonds, ions may be necessary for binding, the receptor may change shape (“induced fit”) and rotatable bonds in the ligand and receptor might have to be arranged before finally the ligand is said to be “bound” to the receptor. A detailed summary of those steps is given in Figure 7. The

pharmacodynamics (on-off) rates of the ligand are also of importance therefore the stability of the complex strongly influences the bioactivity observed.

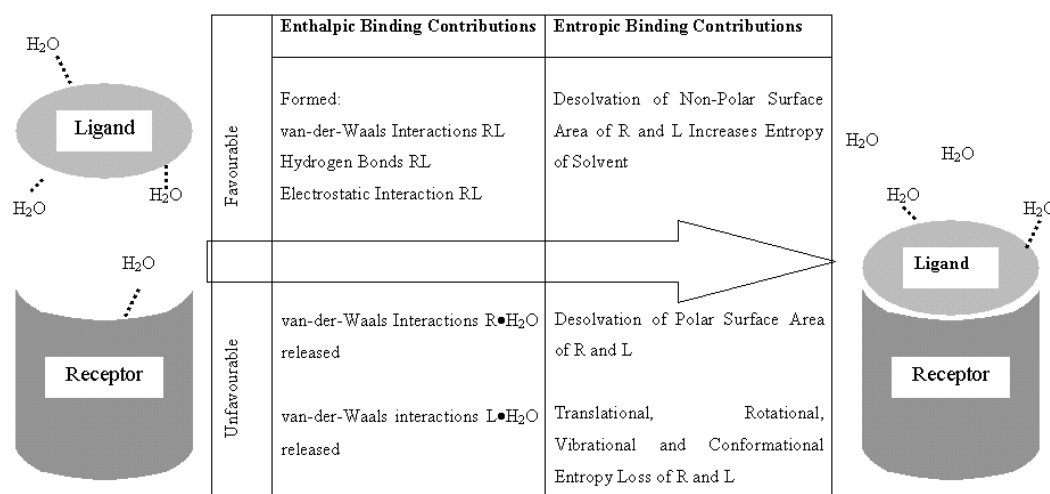


Figure 7. Enthalpic and entropic contributions to ligand (L) – receptor (R) binding. It is commonly assumed that non-polar surface areas contribute to stability of the ligand-receptor complex through desolvation energies while polar surface areas are important for selectivity.

It is commonly accepted that the energy gained during ligand-receptor complex formation by removing water from lipophilic surfaces is largely responsible for the overall stability of the complex formed¹⁷⁵. We commonly observe that, in general, larger ligands have greater affinities (and medicinal chemistry often chases activity by increasing lipophilicity, while unfortunately reducing the solubility and deteriorating pharmacokinetic profiles). This tends to allow more hydrogen bonding within the solvent (water), thus ‘squeezing out’ the ligand from solvent onto the more hydrophobic receptor surface. (While this explanation of the ‘hydrophobic effect’ emphasizes the enthalpic aspects, an entropic explanation can also be attempted, involving the different degrees of rotational (and vibrational) degrees of freedom if water is adjacent to the protein surface *versus* in bulk state.) In contrast, hydrophilic (charged or polar) areas are seen as having adverse effects on overall stability of the complex, because the surrounding arrangement of water molecules can interrupt the ligand/protein interaction by solvating the ligand and protein. Hydrophilic contacts (in particular hydrogen bonding interactions) are thought to possess discriminatory properties between similar binding pockets, thus contributing to selectivity^{176,177}. In many successfully applied molecular similarity methods, such as Comparative Molecular Field Analysis (CoMFA, GRID), steric and electrostatic fields are

employed. Steric effects (where dispersion is ignored) increase with the proximity of interacting groups. Electrostatic effects increase with the proximity of like charged fragments. However, this is an approximation (sometimes useful) of the true solvated situation. As we outline in the following paragraphs, electrostatic fields may not always be appropriate for generating descriptors for molecular similarity calculations and this observation opens up possibilities for improvements in this area.

Examination of the literature does not give a conclusive account of the relative importance of steric and electrostatic contributions to the predictivity of Comparative Molecular Field Analysis (CoMFA)³² models. This may also be due to the different nature of problems the method is applied to. An overview of the relative importance of both computed fields in the literature is given in Table 2.

Table 2. Importance of steric and electrostatic field components in applications of Comparative Molecular Field Analysis (CoMFA)³². Interestingly, electrostatic and steric effects often contribute in significantly varying degrees to CoMFA models.

Steric > Electrostatic	Steric ~ Electrostatic	Electrostatic > Steric
Selection of the most predictive conformation of Adenosine A2A receptor agonists ¹⁷⁸	Alkylamides as inducers of human leukemia cell differentiation ¹⁷⁹	Affinity of dyes for cellulose fiber ¹⁸⁰
Dopamine D4 Receptor Antagonists ¹⁸¹	SAR of antifungal pyrrole derivatives ¹⁸²	
	Flavonoids Binding at Benzodiazepine Site in GABAA Receptors ¹⁸³	
	Cytotoxicity of substituted acridines against HCT-8 cell line relative to mouse leukaemia L1210 cells ¹⁸⁴	

In order to examine predictivity using electrostatic information, Chau and Dean studied electrostatic complementarity of 34 ligand-receptor complexes¹⁸⁵. Calculating correlation coefficients of van der Waals surface points, they found significant correlation between surface complementarity and binding affinity in all but eight cases - but with a negative slope. This indicates that electrostatic complementarity is far from being sufficient for binding, probably due to desolvation effects of both receptor and ligand. Indeed, there is

no reason to believe that electrostatics should have only a positive or negative contribution. In a set of molecules, individual sites can in fact be correlated in terms of their electrostatic changes (e.g. partial charge on a fragment can increase across a series) but seemingly chaotic in their contribution to binding (binding goes up or down irrespective of charge). Many methods, such as CoMFA, rely on discovering relationships (in this case linear, using PLS) between molecular property and binding affinity. But what happens if the property increases then decreases and affinity increases? The property is discarded in the model. Indeed, in docking studies, it was found beneficial to include desolvation information that was dependant on the pairwise fragment interactions between hydrogen bonding groups. Affinity between fragments in the ligand and protein can be attractive or repulsive depending on the pair involved. This implies that electrostatics alone can often be insufficient to account for changes in affinity. The perturbations introduced by the receptor on the ligand in terms of changes in desolvation energy can be fundamental to molecular recognition¹⁸⁶.

Klebe and Abraham¹⁸⁷ discusses the influence of enthalpic and entropic factors on binding. Building CoMFA models for Renin inhibitors, he concludes that only binding enthalpies and not free energies can be predicted from the models. Since the difference between binding enthalpy and binding free energy is explained by the entropy term ($\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$), this is in agreement with the importance of the desolvation energy caused by entropy changes of the solvent and ligand upon binding. In the context of desolvation enthalpies and entropies and ligand entropies, the “totally unexpected”¹⁸⁷ observations are explained.

Where transport properties like partitioning are important for activity, the use of steric and electrostatic field based methods may not be as useful as single parameters like logP which in these instances may have more predictive value. This is understandable, since logP introduces information about the hydrophobic character of a molecule, which is not captured by both steric and electrostatic fields. Still, hydrophobic parts of the molecule possess important information for overall affinity since their binding contribution is often positive. Soon after publication of CoMFA³², a third, hydrophobic field was introduced¹⁸⁸. Very little advantage (in this context) compared to the original method has been found, which may be attributable to a number of reasons. Some of the information (variation) in the hydrophobic field may already implicitly be present in steric and electrostatic fields. Also, variable selection using partial least squares penalizes the increased number of variables. More likely, it seems that similar problems from heterogeneity in the local

binding contributions to those found with electrostatics are introducing non-linear interaction trends that cannot be handled by a linear least squares method.

However, one advantage of the PLS coefficients resulting from hydrophobic fields is that they are easy to interpret¹⁸⁹ and are very useful in identifying similar substituents having similar properties when observed in the context of the grid computations and some later studies have suggested that the cross-validated correlation coefficients obtained from these fields are superior to those from steric and electrostatic fields, e.g. as is reported for multidrug resistance reversal¹⁹⁰.

The recently published HINT (Hydrophobic INTERaction) force field attempts to include hydrophobic information, derived from logP data¹⁸⁹. Other molecular descriptors, such as GRIND³⁵ or CoMSIA^{105,106} also make use of hydrophobic fields, as does molecular lipophilic potential (MLP)¹⁹¹.

To summarize, most commonly employed molecular similarity methods, such as in the original CoMFA³², in which ligands are compared to each other, often appear to work well without taking any of the effects listed in Figure 7 into account. However, electrostatic fields alone conceptually do not capture relevant information because they neglect desolvation energies. Not considering solvation and desolvation effects may lead to inappropriately fitted models, which, even if they show some correlation with experimental data, can in the worst case only be seen as random fits. In ligand/protein binding, the enthalpic and entropic influence of solvation and desolvation is crucial for the assessment of molecular similarity. Better understanding of this phenomenon and the use of statistical methods that could take this into account could offer great benefits and improve methods like CoMFA considerably.

Reviews of forces influencing receptor-ligand formation are given by Bohm and Klebe¹⁹², Gohlke and Klebe¹⁹³ and Brooijmans and Kuntz¹⁷⁵.

6.7 Local similarity requires non-linear models. or: why linear regression techniques are not the method of choice.

There have been many analyses using e.g. feature selection followed by multiple linear regression using computed properties to determine some bioactivity of interest. Using force fields to calculate individual contributions to the binding free energy for molecular features, the COMBINE analysis method¹⁹⁴ uses interaction energies of individual groups to predict interaction free energies in a linearly-additive fashion. The underlying assumption is that there is some linear and additive relationship between the contributions

each of the properties makes to bioactivity. These calculations can be carried out on the molecular graph without the availability of 3D information or can also include properties derived from the 3D structure. Given that there are often local neighbourhoods of similar activities that are discontinuous, may these regression methods simply be connecting neighbourhoods (means of clusters of similar active molecules)? The criticism of linear regression in QSAR studies has already found consideration in recent research. Multiple linear regression and smoothed splines have been compared, and the non-linear smoothed splines have been found to be superior¹⁹⁵. One has to be careful not to draw too far-reaching conclusions because a single – and relatively small – dataset was used in the study. Also local linear and non-linear models, which are able to take different modes of action into account, were examined by Ren¹⁹⁶. Locally weighted regression scatter plot smoothing (LOESS), multivariate adaptive regression splines (MARS), neural networks (NN), and projection pursuit regression (PPR) were applied in this case to toxicity prediction. K-nearest neighbour QSAR follows directly from the idea of locality of bioactivity space and considerable improvements in predictivity were recently achieved on some QSAR datasets^{1,197}.

In 3D studies using regression methods, alignment and the discontinuous properties of substitutions of key interacting groups can have a similar effect. Dissimilar structures may still exert the same effect on a receptor. Those examples of functional equivalence without structural similarity are well known¹⁹⁸. The concept of bioisosterism holds surprises even for experienced medicinal chemists, recently for example metallocene-derived ligands of GPCRs with subnanomolar binding affinity were reported¹⁹⁹. Linear models also reduce the binding model to a single binding site and to a single binding mode, an assumption that is often far from being true. Slight changes in molecular structure may cause a completely different binding mode, such as that of the DHFR inhibitor methotrexate, compared to the binding mode of the natural substrate²⁰⁰. In this case, the binding orientation was completely inverted following slight changes of molecular structure. To address one aspect of this, multimode ligand binding was implemented into CoMFA²⁰¹ and appears to give better results than standard CoMFA.

While it may be sensible to apply linear methods to the prediction of physicochemical properties¹²², there is conceptually a problem when applying linear regression to relationships between structure and biological activity (in the ligand/receptor scenario). While a single linear model may be preferable to some researchers, it is not necessarily the case that such a model exists (in particular in the presence of multiple binding modes).

If multiple binding modes are encompassed in a single linear model then a predictive outcome may only be due to binding contributions that stem from similar molecular features in both binding modes – and there is no reason to assume that this assumption should be true in general. This problem is illustrated in Figure 8.

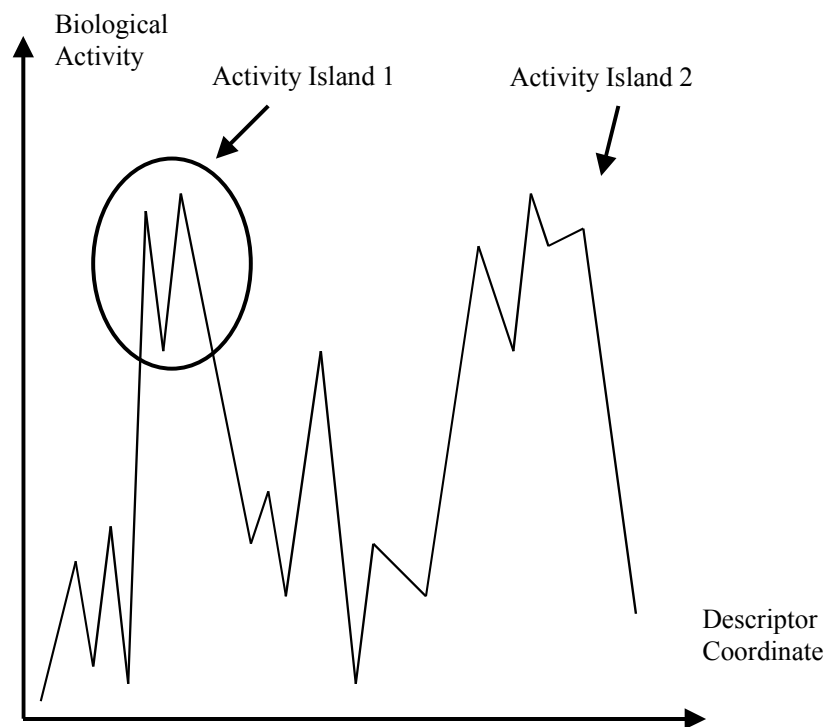


Figure 8. Variations of a molecular property (here biological activity), depending on the positioning of the structures in property space. Linear regression, which may be appropriate in case of physicochemical properties such as logP, is certainly not appropriate to model this relationship between structure and activity. (Non-linear) Methods based on clustering around activity neighbourhoods should perform better.

Different structures are placed at different positions in chemical space (here illustrated using a one-dimensional descriptor axis). Bioactivity does not depend linearly on the position in descriptor space. Instead, local “activity islands” are encountered, which can only be modelled using non-linear models or recursive partitioning. Regression through these local islands results in a modest regression coefficient and occasional outliers – a situation commonly found in analyses of this type. The removal of outliers (with various ‘reasonable’ excuses) is a common feature of many QSAR analyses that was recently discussed in-depth²⁰²⁻²⁰⁴.

6.8 Substructural Analysis, Binary Kernel Discrimination and the Naïve Bayes Classifier

A Short History of Substructural Analysis

Substructural Analysis attempts to explain observed molecular properties (for example bioactivities) by the presence and absence of molecular features (substructures). This field, conceptually still very similar to the approaches employed today, was opened by Cramer *et al.* in 1974⁶⁸ and the work shall briefly be summarized here. Of a dataset of about 700 compounds which were evaluated with respect to their bioactivity (“reduction of hind paw volume”) 189 compounds were found to be active (significantly reduced hind paw volume). 492 substructural fragments were identified in the whole dataset and a table containing the frequency of fragments in the active and inactive dataset was constructed which was used to make predictions about the activity of an additional 700 compounds. Both internally and on the new dataset it could conclusively be shown that molecules showing particular substructures were more active than molecules without those substructures. The problem of very infrequent fragments was clearly perceived, namely that - in the most extreme case, given one molecule containing the substructure - unique fragments can only be “strongly correlated with activity”, since “all” molecules containing this fragment are active, or “strongly anti-correlated with activity”, since “none” of the molecules is active. Also the concepts of internal cross-validation as well as the application to a new, external validation set were already applied. While atom typing, fragment definition and validation of the approach using larger and more diverse datasets are comparatively recent developments, the conceptual basis of the approach is still employed today.

The weighting of features which are present in the individual molecules can be performed using one of a different number of “weights”²⁰⁵. The two principles which can be followed here are the “independence assumption” and the “ordering principle”, leading to a total of four possible weighting schemes²⁰⁵ (given two possible interpretations of each of the principles). The independence assumption decides whether the frequency of features among (in our case) the active molecules should be compared to the frequency of features among only the inactive molecules, or to that in the whole corpus of structures. The ordering principle decides whether attention should be paid only to those features which are *present*, or also to *absent* features. In effect, this question decides whether common absence of features should be perceived as conferring similarity between instances or not. When comparing molecules both of those points need to be decided upon. In our classification model shown later, we compare frequencies of features among active compounds and compare them to those among (assumed) inactive compounds only.

Also, we only include those features which are present in compounds and neglect features which are absent in both structures to be compared.

Besides only comparing individual molecules also information from multiple structures can be extracted and machine learning algorithms can be applied to make predictions about novel compounds. Here two approaches of increasing interest in recent times shall be mentioned, namely Binary Kernel Discrimination and the Naïve Bayesian Classifier.

Machine Learning Methods using Substructural Analysis – Binary Kernel Discrimination (BKD)

Kernel discrimination employs a high-dimensional representation of data points in space (whose shape is defined by a Kernel function) in order to make distance-dependent predictions about the behaviour new compounds in the same space. Since the ideal shape of the transformed space is not given *per se*, the best parameters for this transformation need to be determined empirically. More specifically, the “smoothing parameter”, which determines how far out points influence the space around them, needs to be optimized for each individual classification task. This parameter is conventionally referred to as λ .

Historically the first application of BKD was presented by Aitchison and Aitken²⁰⁶ who translated the use of real-valued applications to binary descriptors in the medicine area.

The application of Binary Kernel Discrimination (BKD) in ligand-based virtual screening started with the work of Harper *et al.*²⁰⁷ who employed it for the detection of inhibitors of Monoamineoxidase (MAO). This dataset is referred to later in this thesis, as are results of the methods presented in this thesis compared to the performance of the BKD classifier. Harper *et al.* suggested the following kernel for two molecules, *i* and *j*, which are represented by binary fingerprints with length *M* whose bit patterns differ in a number of d_{ij} positions:

$$K_{\lambda}(i, j) = \lambda^{M-d_{ij}} (1 - \lambda)^{d_{ij}} .$$

Sensible values for the smoothing parameter, λ , are in the range $0.5 \leq \lambda \leq 1.0$, where a value of 0.5 indicates that all training set molecules have the same influence on the overall score of the molecule whose properties are predicted, while values λ approaching 1 lead to the domination of the score by the largest kernel functions. The total score L_{λ} for each molecule is then calculated by taking the score above, for each individual molecule pair

from query and known active and inactive sets, and combining them *via* the following formula:

$$L_A(j) = \frac{\sum_{i \in Active} K_\lambda(i, j)}{\sum_{i \in Inactive} K_\lambda(i, j)}.$$

While the smoothing parameter λ needs to be optimized for each individual dataset, this step can also be interpreted in the way that the classifier can be *adapted* to the nature of each dataset, *w.r.t.* its diversity and its distinct peculiarities discriminating between the compound classes.

Machine Learning Methods using Substructural Analysis – The Bayes Classifier

The Bayes Classifier predicts likelihoods for class membership of an unseen instance, given conditional probabilities (relative frequencies) of features derived from known instances. In its simplest form the following equation for the Bayes Theorem can be written:

$$P(A | B) = \frac{P(A)P(B | A)}{P(B)}$$

where $P(A)$ is the *a priori probability* of A (the assumed probability of the class before any training instances are seen), $P(B)$ is the a priory probability of B, and $P(B|A)$ is the conditional probability of B being true, given A. It is derived from the following two equations:

$$P(A | B) = \frac{P(A, B)}{P(B)} \quad \text{and}$$

$$P(B | A) = \frac{P(A, B)}{P(A)}$$

which connect the *a priori* probability of one event *via* the conditional probability of a second event to the conditional probability of A, given B, as well as B, given A. By

substituting for $P(A,B)$ in one of the above formulas Bayes Theorem can be obtained. Since in this work the Bayes Classifier is heavily employed for compound classification a related implementation of the algorithm shall be discussed, which is then compared to the approach employed in the current work.

One of the most recent implementations of the Bayes Classifier is given by Scitegic²⁰⁸ in their PipelinePilot cheminformatics suite²⁰⁹. The major difference between the original Bayes Classifier and the approach employed by Scitegic is the Laplacian correction which tries to alleviate the influence of very rare features which would, if not corrected, determine the resulting probabilities to a great extent. More precisely, the Laplacian correction chosen²¹⁰ amounts to

$$P_{Laplace}(Active | F) = \frac{A + P(Active) \cdot K}{B + K}$$

where $P_{Laplace}(Active|F)$ is the Laplacian-corrected probability that a compound showing feature F belongs to the class of active compounds, A is the number of active samples out of which are found to be active, out of B total samples. $P(Active)$ is the *a priori* probability of compounds being active. K is the actual Laplace correction and represents the number of additional “virtual” samples of this individual feature with $K=1/P(Active)$ amounting to the standard Laplacian correction. Thus, for very infrequent (small A) activity-conferring features the Laplacian-corrected score is increased and for very frequent (large A) activity-conferring features the Laplacian-score is decreased, compared to the uncorrected score. This represents a degree of “caution” when gauging new molecules, where extreme probabilities are corrected to have a less severe influence.

In our implementation of the Bayes Classifier, we did not implement a standard Laplacian correction but took a different approach. The influence of very frequent and very rare features was in principle treated like in the original Bayes Classifier, paying respect to the observation that also small probabilities can be significant. In order not to let our classification be overly dominated by over- and underconfident (very large and very small) probabilities, we employed a feature selection step prior to classification. This information-gain feature selection step was able to eliminate non-significant features from the set and has the additional advantage that features “characteristic” for a given activity class could be visualized and interpreted. In addition to correcting for extreme probabilities also zero-probabilities needed to be accounted for. Here we were using

several different treatments of zero-probabilities, among them the $1/m$ -correction (multiplication with class prior probability), multiplication with a fixed constant factor and with a larger punishment factor, $1/2m$. To summarize, this is, on the classification side, the major difference between the two approaches: Where Scitegic uses Laplacian correction, we use information-gain feature selection and a dataset-size dependent factor for non-existent features (zero-probabilities).

In addition to the classification method it should briefly be mentioned that also the descriptors differ slightly from each other. While both are based on circular substructures, atom typing is different; where we use force field (mol2) atom types, Scitegic offers the choice between elemental type (ECFP) fingerprints and functional class (FCFP) fingerprints which assign interaction abilities to the individual atoms (e.g. hydrogen bond donors/acceptors, charge moieties, etc.). Also, Scitegic fingerprints employ features *below* a given diameter for classification (e.g. ECFP₆ employs fragments 6 bonds *and less* for classification), while in the work here only fragments matching exactly a given diameter were used (except necessarily in cases of terminal atoms).

Comparison

Binary Kernel Discrimination and the Naïve Bayes Classifier are both able to incorporate knowledge about multiple active compounds and several publications on the topic appeared recently^{31,211,212}. Applications of the BKD method include its application to a large dataset compiled from the MDDR database³¹. It was found that Binary Kernel Discrimination was indeed a very powerful way to combine knowledge about multiple compounds, showing similar performance than the Bayes Classifier²¹³. The comparison of both approaches is also a topic of this work which is shown in detail in the following chapters.

6.9 Other applications of machine learning methods

Machine learning methods attempt to generate rules or models that cluster similar compounds together, usually to predict the properties of other compounds *via* application of the clustering method or model. Different methods have different advantages. For example, artificial neural networks (ANNs)²¹⁴ are able to model flexibly a relationship between input and output variables, but this flexibility can also be problematic, in particular in the case of underdetermined systems. Inductive logic programming (ILP)²¹⁵ has the advantages that human-understandable rules can be derived and that new

relationships can be inferred, but has problems with noisy data and a potentially very large hypothesis space. Using ILP on different data sets, it was found to perform as least as well as other approaches²¹⁶, although direct comparison is difficult.

Neural Networks have many applications in compound clustering; for example recently ANNs have been applied to profiling of GPCR-active compounds²¹⁷. The “black box” aspect and the tendency to over fit are main criticisms of these methodologies.

In recent years, support vector machines (SVMs)^{218,219} have been employed to molecular similarity problems. SVMs try to maximize the separation boundary of instances from different classes and are sometimes faster to train than artificial neural networks^{218,219}. Some applications of SVMs are summarized below.

When compared to different types of artificial neural networks, radial basis function networks and C5.0 decision trees, SVMs were found to perform considerably better on a dataset of dihydrofolate reductase inhibitors²²⁰. In addition, training was fastest of the methods tried. Applied to a drug/nondrug classification problem, SVMs also outperform ANNs slightly⁵¹. More recently and allowing for more freedom in kernel parameters significant improvement of SVM classification over ANN results was achieved on the same dataset²²¹. This result is consistent and does not depend on descriptors or size of the datasets. In a similar application predicting drug-likeness and agrochemical-likeness²²², SVMs consistently outperform ANNs. An application of SVMs for prediction of ADME properties²²³ has also been performed, using blood-brain barrier penetration, bioavailability and protein binding datasets. For blood-brain barrier penetration and protein binding SVMs with RBF and Quadratic Kernels, respectively, performed best whereas on the bioavailability data set ANNs were of superior performance.

A slightly different application used SVMs in combination with active learning²²⁴. Active learning uses knowledge obtained about formerly untested compounds. A set of active compounds is known from which a model is built. This model is used for screening of a library. Information about the tested compounds in every screening step is then fed back into the model. Active learning has also been combined with other machine learning approaches, such as k-nearest neighbour methods and C4.5 and a simple OR classifier²²⁵. The transductive OR classifier performs best and continuously selects meaningful features. The performance of other methods deteriorates if more and more features are selected.

Other applications of machine learning methods in cheminformatics applications of similarity comprise Binary Kernel Discrimination^{206,207} and the Naïve Bayesian

Classifier^{79,80,138}. Binary Kernel Discrimination in combination with atom pairs and topological torsion descriptors gives robust results (robust to noise) and slightly outperforms neural networks. The Naïve Bayesian Classifier in combination with Atom Environment descriptors is able to accommodate knowledge from multiple active compounds and in test sets examined, outperformed other commonly used methods which are based on both 2-dimensional and 3-dimensional structure^{79,80}. Both descriptors and classification methods in the work of the Naïve Bayesian Classifier are closely related to algorithms published by Scitegic and earlier work on Binary QSAR²²⁶.

6.10 Conclusions and outlook

Molecular similarity is extensively and successfully used in the drug discovery context often to compare molecules in the absence of other mechanistic information (a partial exception is the docking applications described above). Most importantly, similarity has a context. One has to be aware that similarity defined on molecules alone in the absence of the medium in which they act is an incomplete description so great care has to be taken to use descriptors that are appropriate. Back-projectable descriptors possess better interpretability and will probably have more widespread use in the future. Binary bit strings in combination with similarity coefficients possess preferences with respect to bit density (and thus size of the molecule) and combinatorial preferences and one should be aware of these preferences when applying similarity methods. Applications of machine learning methods in computer-aided molecular design will certainly gain importance in the future particularly with the incorporation of heuristics that improve performance.

The discontinuous nature of biological effects such as ligand receptor binding means that linear regression techniques are only appropriate for QSAR and related applications if a linear relationship between feature space and activity exists. In general it is often more appropriate to use nonparametric or non-linear regression techniques. The example of electrostatic effects and their discontinuous relationship with solvation energies is an example.

7 Fragment-based similarity searching (MOLPRINT 2D)

As outlined in detail in the introduction, the comparison of molecules consists of the representation of structures in an abstract form (the calculation of ‘descriptors’), optional feature selection and the assignment of a similarity value to either pairs of individual structures or of a model score (which may have been built using multiple structures) to an individual structure.

In this chapter, we describe the utilization of circular fingerprints, information-gain based feature selection and the Naïve Bayesian Classifier for molecular similarity searching.

7.1 Material and methods

a) Descriptor generation / molecular representation

We use atom environments^{79,80} as a molecular representation. (‘Atom environments’ are synonymous with the term ‘MOLPRINT [2D] descriptors’, while the whole sequence of atom environments, information-gain based feature selection and the Bayes Classifier is referred to by the name ‘MOLPRINT 2D [method]’). Atom environments are similar to Signature Molecular Descriptors^{76,227}. They resemble augmented atoms²²⁸; Scitegic Extended Connectivity Fingerprints (ECFP)²⁰⁸ are constructed in a similar fashion with the major difference being the atom type definition used (where we employ Sybyl mol2 atom types). 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 9):

1. Sybyl atom types²²⁹ 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.
2. A count vector is constructed with the vector elements being counts of atom types at a given distance from the central atom. Every heavy 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.

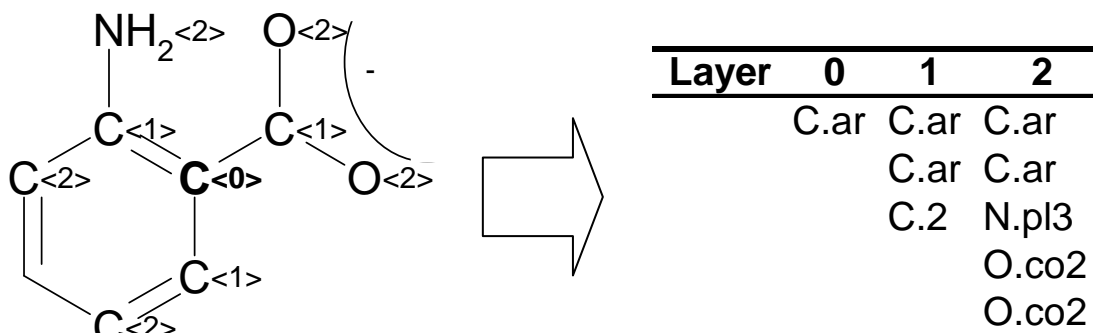


Figure 9. Illustration of descriptor generation step, applied to an aromatic carbon atom. The distances (“layers”) from the central atom are given in brackets. In the first step, Sybyl mol2 atom types are assigned to all atoms in the molecule. In the second step, count vectors from the central atom (here 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.

Descriptors are stored as binary presence / absence features for each molecule. Since they are calculated for every heavy atom of the structure, as many descriptors are calculated as there are heavy atoms present in the structure.

b) Feature Selection

The information content of individual atom environments was computed using the information gain measure of Quinlan²³⁰. 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

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

Where

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

S is the information entropy; S_v is the information entropy in data subset v ; $|D|$ is the total number of data points; $|D_v|$ is the number of data points in subset v and p is the probability that a randomly selected molecule of the whole data set (or subset in case of D_v) belongs to each of the defined classes (equivalent to the relative size of data subsets).

c) Classification

A Naïve Bayesian Classifier²³¹ was employed as a classification tool. The Naïve Bayesian Classifier provides a simple yet surprisingly accurate machine-learning tool²¹. Trained with a given data set which consists of known feature vectors (F) and their associated known classes (CL), a Bayesian Classifier predicts the class that a new feature vector belongs to as the one with the highest probability of $P(CL_v | F)$ which is given by

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

Where

$P(CL_v)$: probability of class v

$P(F)$: feature vector probability and

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

v : class.

In the Naïve Bayesian Classifier we assume that

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

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

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

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

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

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). If a given feature from a new molecule is not present in one of the data subsets CL_1 or CL_2 the probability of class membership for that class would drop to zero immediately. This would cause problems in particular for the denominator in formula 2 as the probability ratio becomes infinite. We use a Laplacian correction to avoid this problem. Features which are not present in data subset v are assumed to be present in a data set of twice the size of the larger of the two subsets. The term $P(f_i | CL_v)$ in this case assumes the value $\frac{1}{2 | D_v |}$. In addition the influence of

different Laplacian corrections has been explored (see calculations section for details).

The underlying assumption of the “Naïve” Bayes Classifier is the independence of features, although it appears to perform surprisingly effectively where features are not strictly independent^{231,232}. Because descriptors calculated from adjacent atoms are often highly correlated, this tolerance towards dependent features is important for our method.

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.

d) Compilation of dataset and preprocessing

As a first dataset for the evaluation of the algorithm, 957 ligands extracted from the MDDR database⁸⁵ were used¹³² (from now on referred to as Dataset A). 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 belong to any of these activity classes.

The method has also been applied to a recently published dataset³¹ derived from the MDDR⁸⁵ (Dataset B). This dataset comprises more than 100,000 structures, which includes 11 sets of active structures. The active datasets range in size from 349 to 1236

structures and are currently one of the largest reference datasets for this type of comparison available. (The precise dataset sizes are given in Table 3.) In a recent publication, similarity searching results using Unity fingerprints and various data fusion methods as well as Binary Kernel Discrimination were published on this dataset³¹ which makes it a comprehensive benchmark for any new method with respect to size and diversity of the structures.

Table 3. Activity classes, MDDR activity IDs and sizes of active datasets derived from the MDDR

Activity Name	MDDR Activity ID	Dataset Size
5HT3 Antagonists	06233	752
5HT1A Agonists	06235	827
5HT Reuptake Inhibitors	06245	359
D2 Antagonists	07701	395
Renin Inhibitors	31420	1130
Angiotensin II AT1 Antagonists	31432	943
Thrombin Inhibitors	37110	803
Substance P Inhibitors	42731	1246
HIV Protease Inhibitors	71523	750
Cyclooxygenase Inhibitors	78331	636
Protein Kinase C Inhibitors	78374	452

From the 102,535 structures of the original dataset, 102,524 could be retrieved from our local MDDR database. Conversion of the dataset to mol2 format gave 102,513 valid structures. This corresponds to 99.98% of the original dataset. All of the structures not retrieved belong to the inactive dataset. All structures in mol2 format could be converted to MOLPRINT 2D fingerprints.

This dataset spans a variety of activities as well as a very large number of compounds with defined endpoints (some of which are however ambiguous) which provides a useful benchmark for a similarity searching method. One has to be aware of the occurrence of close analogues though which favours 2D methods and of the fact that the MDDR does not include explicit information about inactivity of compounds. (Also see the last chapter of this thesis for a more detailed discussion of this point.)

The method was further applied to a monoaminooxidase (MAO) data set^{7,207,233} which comprises 1650 structures (Dataset C). This dataset was previously used as a test set for the investigation of the Binary Kernel Discrimination method²⁰⁷ applied to chemical similarity searching. In this paper, Binary Kernel Discrimination was used in connection with topological torsion and atom pair descriptors. Of the total number of compounds, 1360 were inactive and 290 were active. In the dataset, activities were given in three categories, 1, 2 and 3, which were merged as in the Binary Kernel Discrimination application to give one dataset containing the active molecules.

e) Calculations

Structures were downloaded in SDF format and converted to Sybyl mol2 format using OpenBabel²³⁴ 1.100.1 with the `-d` option to delete hydrogen atoms and default mol2 atom typing. Descriptors were then calculated directly from mol2 files.

Two separate validations of the method presented here were performed on Dataset A. 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. 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. (Note that 100 features corresponds to all features present in the active set, given the number of query structures.) 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 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 into 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 on Dataset A and for ten randomly selected compounds of each of the five classes of active compounds its ten nearest neighbours were calculated. 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. An individual hit rate was calculated for each compound based on the number of molecules within its ten nearest neighbours, which belong to the same activity class as the query compound. The nearest neighbour protocol of Briem¹³² has been followed in this validation to make it easy to compare the performance of our algorithm with commonly used methods.

For all calculations using the large MDDR dataset (Dataset B), the performance measure was the fraction of active compounds found within the first 5% of the sorted library. Sorting was performed by descending Tanimoto similarity or decreasing probability of compounds being active, if the Naïve Bayesian Classifier was employed.

In the first part of the validation runs, 10 active compounds were selected randomly from each of the 11 classes of active compounds and the Tanimoto coefficient was employed. This similarity measure was used by Hert *et al.*³¹ in combination with Unity fingerprints (results are only reproduced here), so it allows us to compare the descriptors only, Atom Environments vs. Unity Fingerprints, since similarity coefficient as well as the datasets used were identical.

In the next calculation, the Tanimoto coefficient was replaced by a Naïve Bayesian Classifier and information-gain based feature selection was added to the algorithm. 10-

fold selection of active compounds from each of the 11 datasets was performed and retrieval rates of active compounds were calculated as described above. As in the application of the binary kernel discriminator by Hert *et al.*³¹, the number of active compounds was set to 10 and all calculations were repeated selecting 100 inactive compounds for each run. 150, 250, 500 (roughly all features present in the active set) and all features present in the complete set were selected in the information-gain based feature selection step when 10 active and 10 inactive structures were used. 250, 500 and all features were selected using the dataset of 10 active and 100 inactive structures.

While performance of the whole method could up to this point be compared to other methods, the performance of individual components (descriptor, machine learning method) could not be deconvoluted. Thus, is the next step ECFP_4 fingerprints, which were previously employed in a comprehensive evaluation of similarity searching methods²¹¹, were used in combination with the Bayes Classifier presented here for comparison. Results are then put into context to the previously reported performances.

In order to investigate the influence of the treatment of ‘0-probabilities’ on classification performance, the Laplacian correction mentioned above was compared to 1/m correction and the omission of features which were present in only one of the datasets.

For the MAO data set²³⁵ (Dataset C), a 10-fold cross validation with a 50/50 random split of both active and inactive structures was performed. (In the paper introducing the Binary Kernel Discrimination²⁰⁷, 5-fold cross validation was performed. The number of runs was increased due to the minimal computational demand of the algorithm, using only seconds of CPU time). As done by Harper *et al.*²⁰⁷, a smaller training set was also used, consisting of a total of 200 randomly selected compounds. The fraction of active compounds found was analyzed depending on the fraction of the sorted library screened. The number of features selected was set to 100, 200 and 500.

7.2 Results and discussion

For the first validation on Dataset A, the influence of the maximum bond distance for creating the 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 4. 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.

Table 4. 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 $m > 2$, $m < \text{max}-2$, 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 artifacts (enrichment factors of 0) in this table.

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 (all)	0.00	0.00	6.52	2.61	0.19	0.08
10, $m > 2$, $m < \text{max}-2$	0.00	0.00	7.33	2.93	3.50	1.40
20, $m > 2$, $m < \text{max}-2$	0.00	0.00	8.58	3.43	3.50	1.40
30, $m > 2$, $m < \text{max}-2$	0.00	0.00	9.20	3.68	3.50	1.40
40, $m > 2$, $m < \text{max}-2$	0.00	0.00	10.28	4.11	3.50	1.40
50, $m > 2$, $m < \text{max}-2$	0.00	0.00	10.13	4.05	3.50	1.40
70, $m > 2$, $m < \text{max}-2$	0.07	0.03	8.99	3.59	3.50	1.40
100, $m > 2$, $m < \text{max}-2$	0.07	0.03	4.89	1.96	0.19	0.08

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 4 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.*^{76,227}. 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.

A visualization of enrichment factors, which depend on the number of selected features is given in Figure 10. In this case, the bond level for descriptor generation, n , has been set to $n = 2$ because it performed best as shown in Table 4. 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.

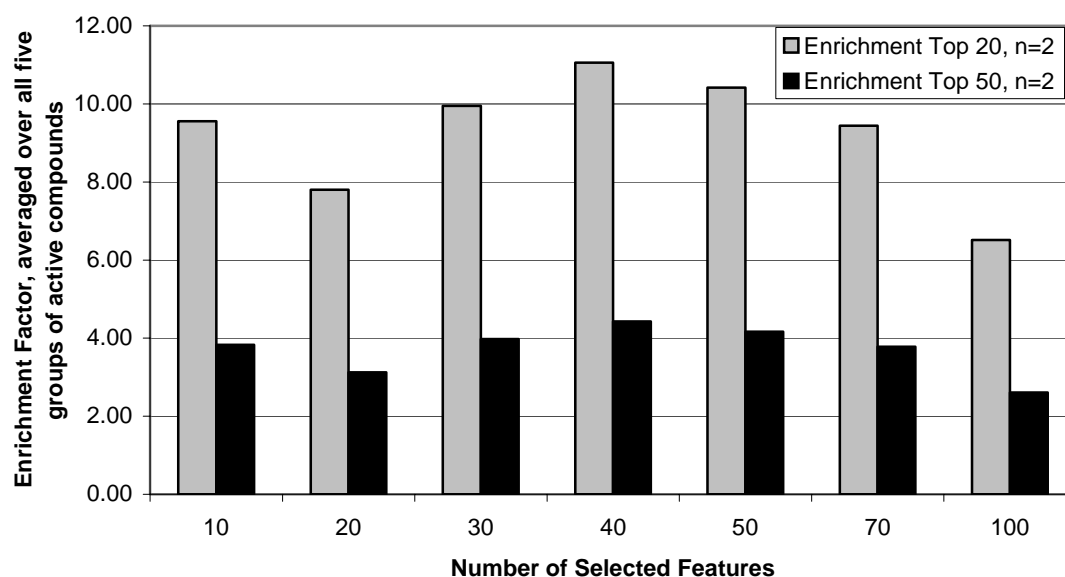


Figure 10. Enrichment factor, averaged over all five groups of active compounds, using atoms up to 2 bonds apart from the central atom to construct the atom environment descriptor and a variable number of selected features for classification. At the given number of molecules, a number of 100 features corresponds to all features present in the active set.

We have found that feature selection has its optimum at a selection of 40 features (i.e., about half the number of features in the active set) 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 5. 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. The performance of the algorithm generally increases if more and more structures are used to generate the query (Table 5), as well as the standard deviation in performance between different sets of query structures decreases. 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.

Table 5. Average hit rates among the ten highest ranked compounds 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.

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

For the best performing method using 40 features with the highest information gain, cumulative recall plots are given in Figure 11. These plots were calculated using 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.

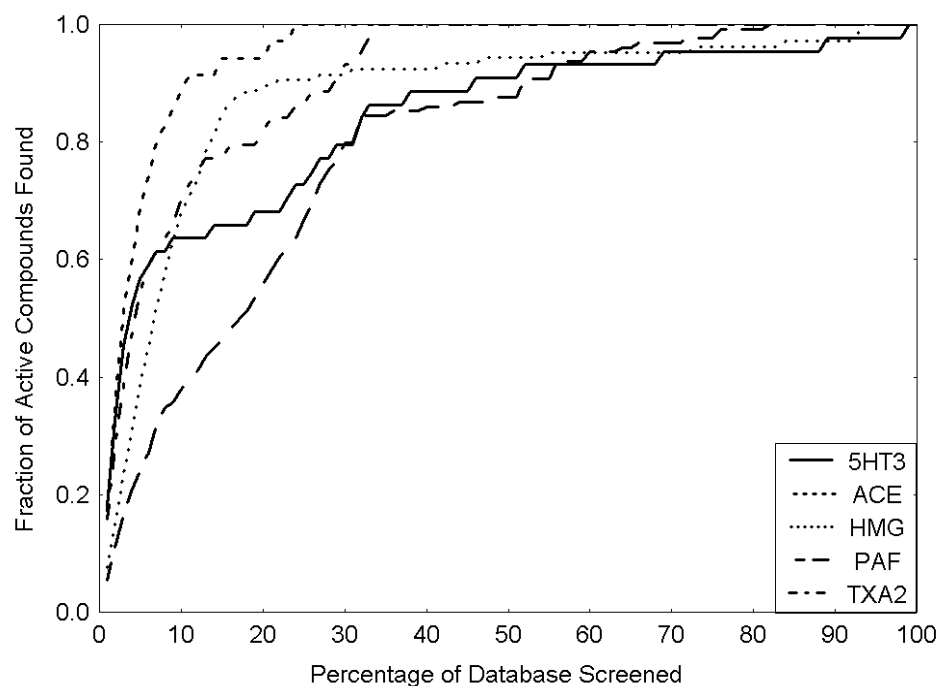


Figure 11. Cumulative recall plot of all five datasets, using atoms up to two bonds apart from the central atom for descriptor generation and 40 features associated with the highest information gain for classification.

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 inhibitors²³⁶ (ACE inhibitors). The design of ACE inhibitors originally followed the hypothesis that ACE had binding site homology with carboxypeptidase-A. A number of interaction sites were proposed based on analog design, shown in Figure 12.

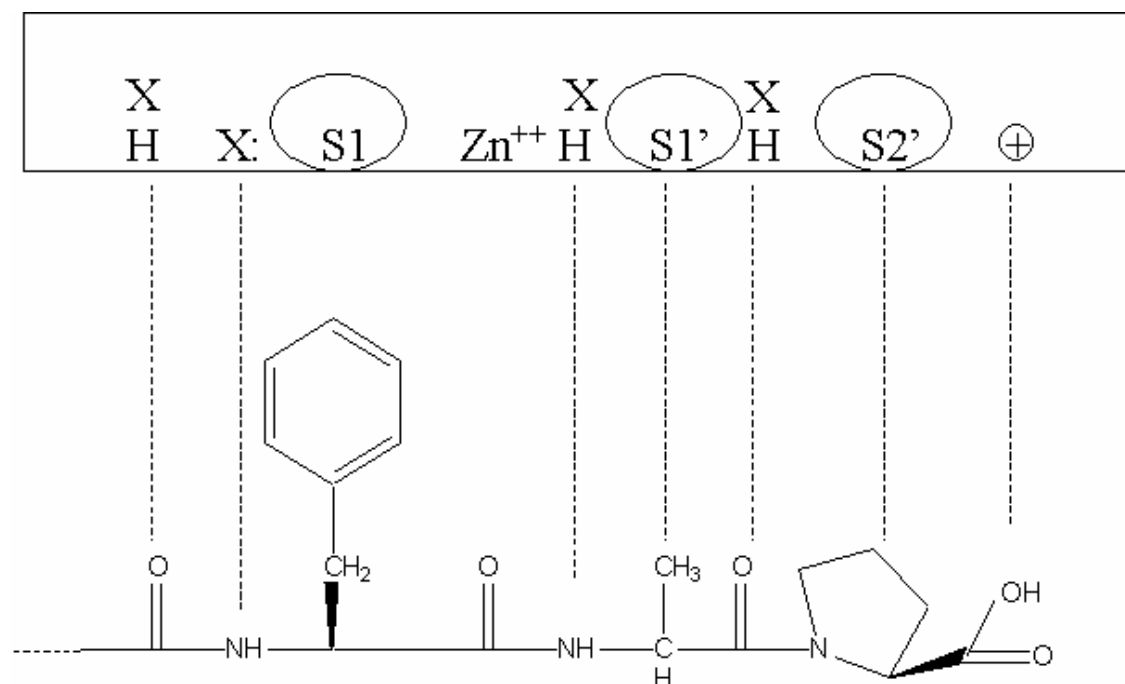


Figure 12. Snake venom peptide analog with putative binding motif to angiotensin used in early compound design²³⁷

A recent crystallographic study²³⁸ of an ACE inhibitor, lisinopril (N2-[(S)-1-carboxy-3-phenylpropyl]-L-lysyl-L-proline), has revealed the binding site interactions in some detail. 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 analogs as being important.

The 5 molecules used to construct the query are shown in Figure 13, the 10 selected features giving the highest information gain are given in Table 6 and the 10 highest ranked structures from the sorted library are shown in Table 7.

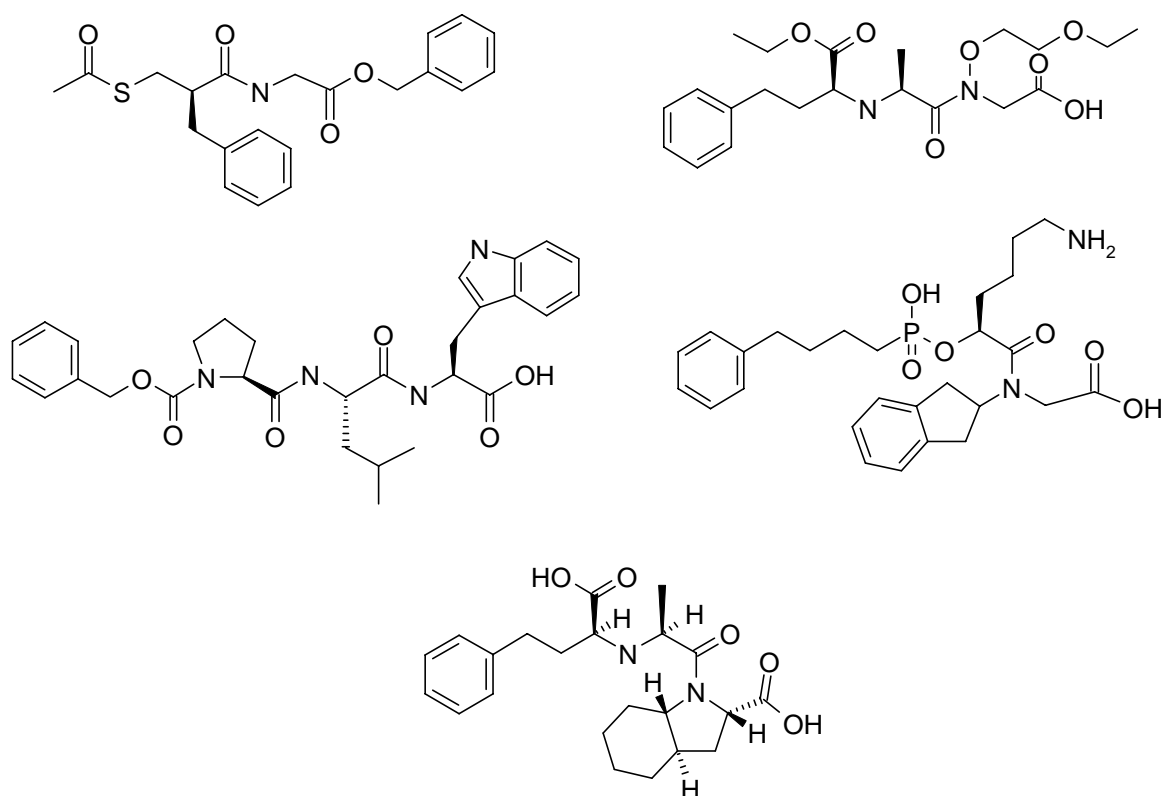
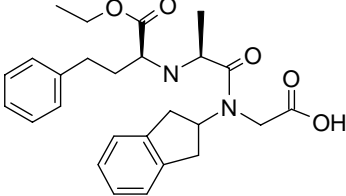
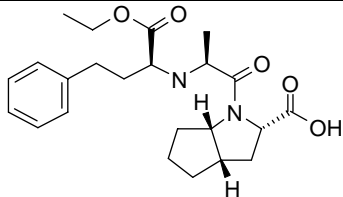
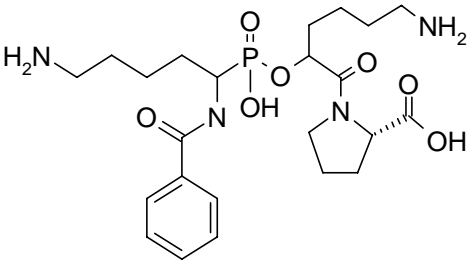
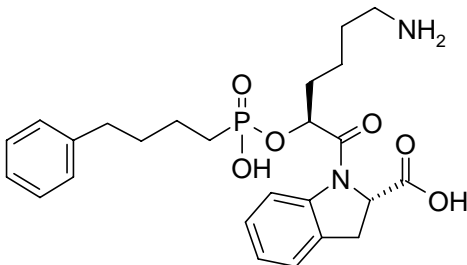
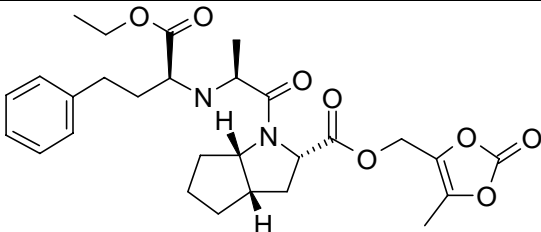


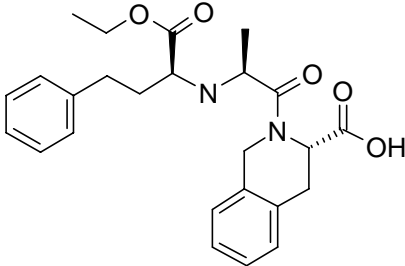
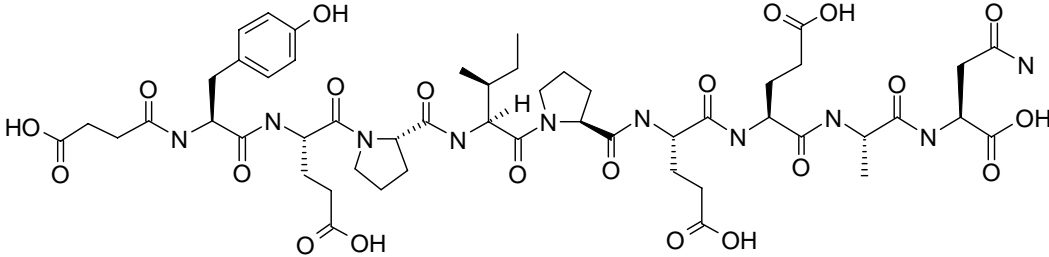
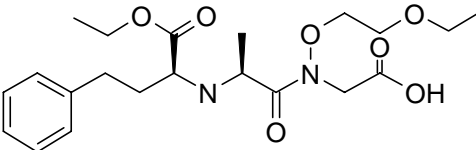
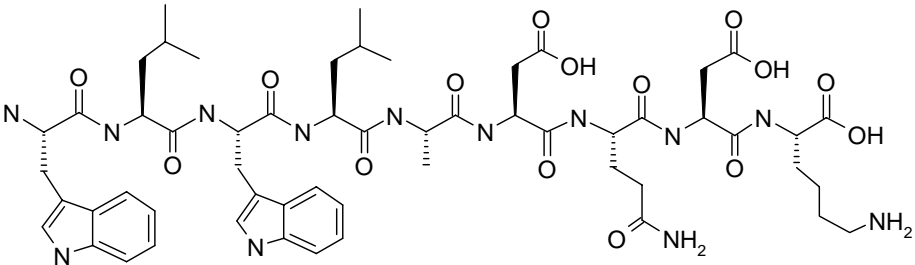
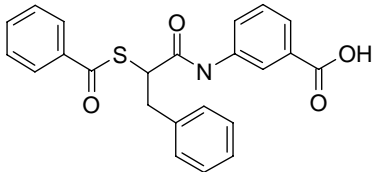
Figure 13. Five active molecules from the dataset of ACE inhibitors, used to construct the query and perform feature selection.

Table 6. Ten features associated with the highest information gains from a sample run using 5 inhibitors from the ACE dataset.

Selected Feature	Information Gain Associated with this Feature	Putative Interaction Site on ACE
	0.0171	S2
	0.0141	Zn ⁺⁺
	0.0127	S2
	0.0118	S1
	0.0114	XH/Zn ⁺⁺
	0.0104	S1
	0.0096	S2/+
	0.0086	S1
	0.0083	S2/+
	0.0083	S2

Table 7. Top 10 ranking molecules of the sorted library. Out of these, seven are active ACE inhibitors and three are inactive molecules in this respect.

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

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

The selected features, given in Table 6, possess carbon, nitrogen as well as oxygen atoms as central atoms. Overall the selection of fragments of ACE inhibitors seems consistent with the binding information deduced crystallographically²³⁸. The five fragments associated with the highest information gain given in Table 6 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 7, 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 analogs 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.

As described in Table 8, 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.

Table 8. 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.

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

The nearest neighbour protocol of Briem 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 Trees⁷¹, ISIS MOLSKEYS²³⁹, Daylight Fingerprints⁸⁶, SYBYL Hologram QSAR Fingerprints⁸⁷ and FLEXSIM-X¹³⁰, FLEXS²⁴⁰ and DOCKSIM¹²⁹ virtual affinity fingerprints. Although the objective of all the methods is the same – the identification of “novel” (in a retrospective sense) active compounds, given a set of known actives – the information contained in the descriptions as well as they way it is exploited differs greatly.

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 MOLSKEYS use 166 predefined two-dimensional fragments for describing a structure. Thus this descriptor belongs to the group of “structural key” descriptors. Daylight Fingerprints are algorithmically generated and describe atom paths of variable length up to a distance of seven bonds: 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. This also explains the name of “hologram” QSAR, since a third (stereochemical) information layer is present. The virtual affinity fingerprints, here comprising Flexsim-X, Flexsim-S and DOCKSIM, describe a molecule by real-valued docking-derived affinities to a panel of (typically in the order of 10) target proteins. They are especially interesting in that the affinity of a ligand to a new target can often be expressed as a linear combination of affinities to known targets²⁴¹. Here it should be noted that they certainly require more computer power for their computation since multiple docking, to each panel protein, is required for each individual ligand. In theory though the outcome should be superior to ligand-based approaches alone, given that target-knowledge is available (though not about the particular target but to a protein panel). 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 14.

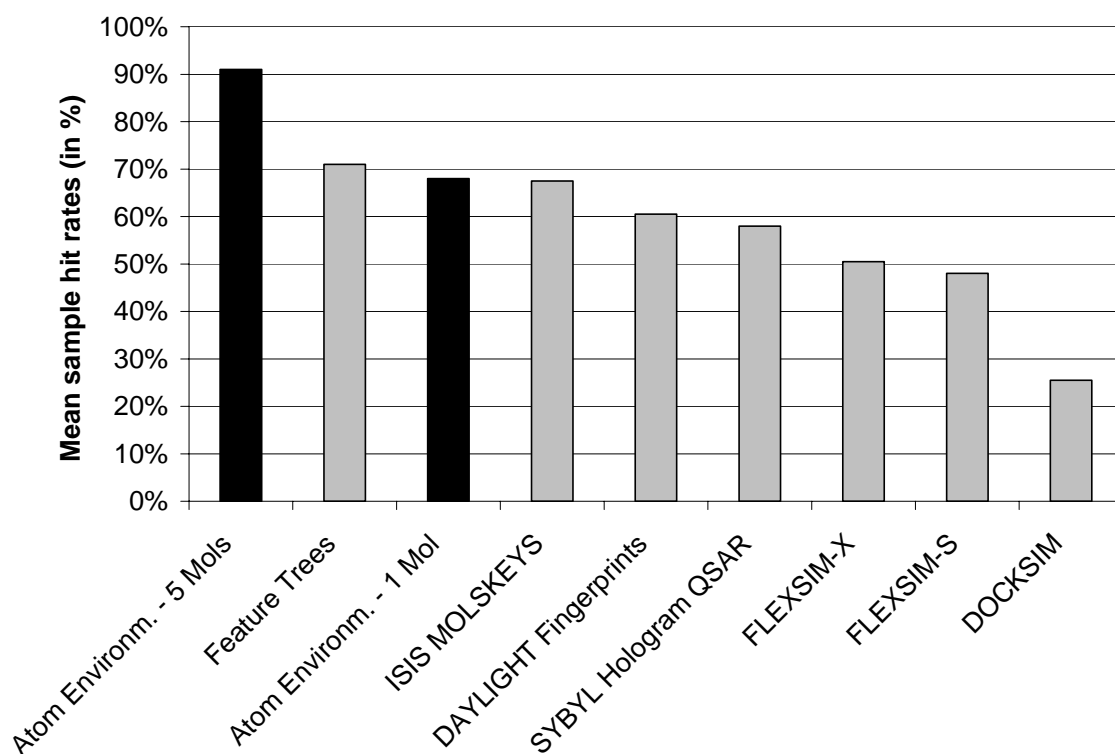


Figure 14. Mean sample hit rates of the Atom Environment approach (black), in comparison to the methods applied by Briem (light grey). The performance of the Atom Environment approach is shown using single queries and randomly selected subsets of five query molecules.

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 Environment 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. Note that all fingerprints, except the ones developed in this study, have been used in combination with the Euclidean Distance instead of the Tanimoto Coefficient, also biasing performance of the descriptors. This is a result of the real-valued nature of the virtual affinity fingerprints which consist of real-valued binding energies. Also, Sybyl Hologram QSAR fragments keep quantitative (integer-valued) counts of fragments which can be used directly for the calculation of Euclidean Distances between compound descriptor representations. An additional more practical consideration was that not all of the above methods were available in-house, which rendered comparison using different similarity coefficients (namely, the Tanimoto Coefficient) not possible.

Using five query structures, the Atom Environment approach achieves a mean sample hit rate of greater than 90%. Although this number is not directly comparable to the other methods it was compared to, it clearly illustrates the advantage of a method of being able to accommodate knowledge from multiple active compounds.

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 up¹³². 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 and Lessel²⁴² give more details about variations in performance among virtual affinity fingerprint based techniques.

The individual hit rates for each group of active compounds from the small dataset derived from the MDDR database are given in Table 9, comparing utilization of the Bayes Classifier and the Tanimoto Coefficient. Note that the utilization of the Bayes Classifier in combination with single molecules is rarely useful in practice since probabilities of features are the essential concept this simple machine learning method utilizes.

Table 9. Hit rates and enrichment factors among the top 10 compounds of the sorted library using the Tanimoto coefficient and using the Naïve Bayesian Classifier, selection of 10 features by information gain. Numbers in parentheses are standard deviations of the mean values.

Performance of the Atom Environment Approach						
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, <i>Using Tanimoto Coefficient</i>						
	7.4	7.8	8.6	7.7	6.6	7.5
	(2.2)	(2.6)	(2.1)	(2.3)	(2.2)	(2.3)
Enrichment Factor	14.8	19.0	7.5	5.5	13.2	9.5
Average Number of Active Compounds Among Top 10 Ranked Compounds, <i>Using Bayes Classifier and 10 Features</i>						
	6.0	8.6	8.1	7.0	6.0	7.1
	(3.2)	(1.8)	(2.8)	(2.2)	(3.0)	(2.6)
Enrichment Factor	12.0	21.0	7.0	5.0	12.0	9.0

Using the Tanimoto coefficient, the average number of active compounds among the top 10 ranked compounds varies from 7.4 (5HT3) and 8.6 (HMG), with an overall average of 7.5. These numbers result in enrichment factors between 5.5 (PAF) and 19.0 (ACE) with an average value of 9.5. These results are slightly better than using the Naïve Bayesian Classifier and single query structures.

The performance of the algorithm presented here is compared to established methods and is shown in Figure 15.

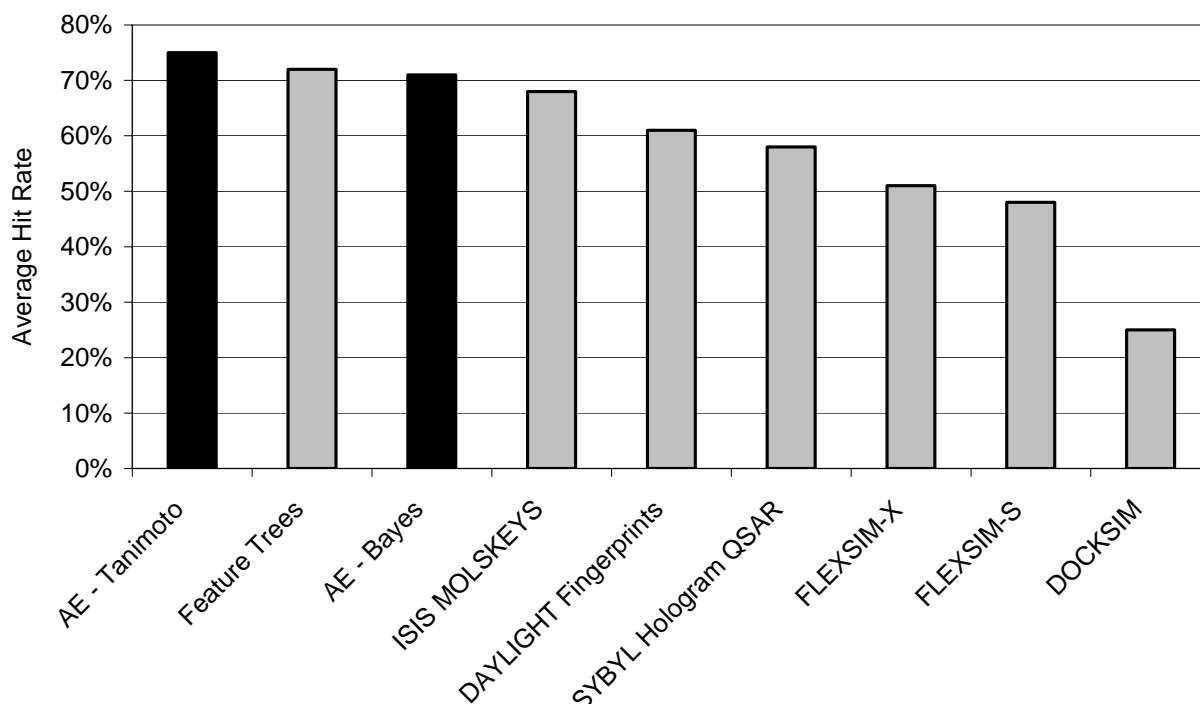


Figure 15. Mean sample hit rates of the Atom Environment (AE) approach (black), in comparison to the methods applied by Briem and Lessel (light grey). The performance of the Atom Environment (AE) approach is on the one hand shown using single queries combined with information-gain based feature selection and on the other hand using the Tanimoto coefficient instead of the Bayesian Classifier.

Shown are mean sample hit rates averaged over all five classes of active compounds. Using one query and information-gain based feature selection 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. If Atom Environments in combination with the Tanimoto Coefficient are used, all commonly employed two-dimensional methods are outperformed slightly.

The influence of the number of active structures used to generate the query, the number of features selected and the feature selection method on search performance is shown in Table 10. Performance at 10, 50, 100 and all selected features is shown in Figure 16.

Table 10. Average hit rates over all five groups of active compounds using different numbers of active compounds for query generation and different feature selection methods. For comparison, performance using all features is given. Numbers in parentheses are standard deviations of the mean values.

Number of Selected Features	Selected by Information Gain (IG) or Relative Frequency (F)	Number of Active Molecules for Query Generation				
		1	2	3	5	10
10	IG	7.1 (2.6)	7.8 (1.8)	8.1 (1.7)	7.7 (2.2)	8.3 (1.2)
10	F	3.4 (2.1)	4.8 (2.4)	5.8 (2.0)	6.0 (2.1)	6.8 (1.4)
20	IG	7.1 (2.5)	7.8 (1.6)	8.2 (1.9)	8.7 (1.1)	8.6 (1.0)
20	F	4.7 (2.6)	6.1 (2.3)	6.4 (2.6)	7.0 (1.8)	8.2 (1.2)
50	IG	5.9 (2.8)	7.7 (1.9)	8.2 (1.4)	8.3 (1.3)	9.1 (0.8)
50	F	3.9 (2.4)	6.0 (2.1)	6.9 (2.0)	7.5 (1.7)	8.4 (1.1)
100	IG	3.4 (2.0)	5.5 (2.0)	7.6 (1.7)	8.2 (1.3)	8.6 (0.9)
100	F	2.3 (1.6)	5.1 (1.7)	6.8 (1.7)	7.4 (1.4)	8.7 (0.8)
All		0.4 (0.0)	0.6 (0.5)	1.2 (0.5)	2.1 (0.8)	5.2 (1.3)

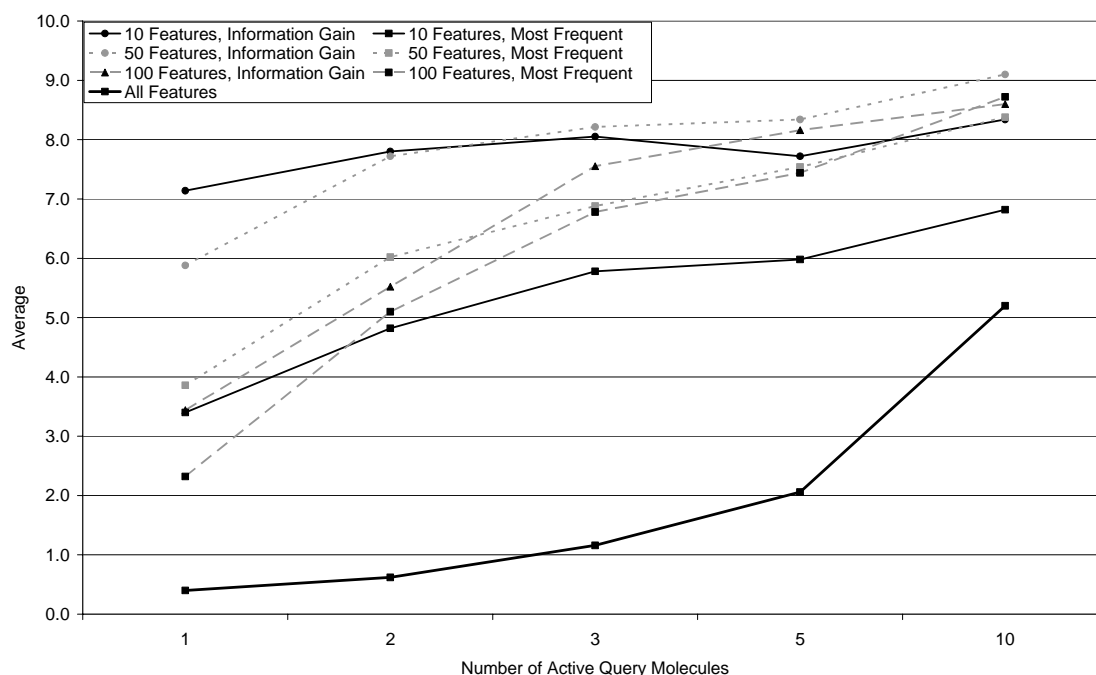


Figure 16. Influence of the number of active query molecules, the number of selected features and the feature selection method (information gain, marked by circles, *vs.* the selection of the most frequent features, marked by square dots) on performance of the similarity searching algorithm. For comparison, results using all features are shown. Information-gain based feature selection performs better than the selection of the most frequent features.

We see a number of consistent trends in Figure 16. Performance generally increases with the number of active molecules used for query generation. If 10 features are selected, performance only marginally depends on the number of active structures. If more and more features are selected, performance increases with the number of active compounds available for training. Information-gain based feature selection generally outperforms selection of the most frequent features as well as using the Bayesian Classifier without any feature selection.

Employment of the Tanimoto similarity measure gives performance slightly superior to all other employed similarity methods (average hit rate of 0.75 over all five classes of active molecules). We can conclude that Atom Environments capture chemically meaningful information for similarity searching of bioactive molecules and that they also perform well in combination with a commonly used (Tanimoto) similarity coefficient.

Most of the variations in performance with the number of active structures and the number and type of features selected (Table 10 and Figure 16) are intuitively explicable.

Performance increases with the number of active structures because meaningful probabilities for features can now be calculated. They are less likely to be random probabilities and more likely to be drawn from the underlying distribution of features of all active compounds. In addition, knowledge about different chemotypes becomes available, depending on the exact query structures used for training.

If more features are selected we see that more active structures have to be available for training in order to achieve best performance. In order to calculate meaningful information gains for a larger and larger number of features, knowledge about relative frequencies of occurrence have to be known for a larger number of features. This is only given if more training structures and thus features are given. If all features are selected, performance increases rapidly with the number of available training structures. This is due to a property inherent in the feature selection step: features have to occur with a certain frequency in order to possess information content relevant to the classification task and high information gain. Thus, the features possessing highest information gain are also present near the top of the list of most frequent features. The set of the most frequent features and the set of the most meaningful features overlap.

Information-gain based feature selection generally outperforms selection of the most frequent features. It is also superior to employing the Bayesian Classifier without any feature selection, in particular with a small number of training structures.

This clearly demonstrates the importance of feature selection for the algorithms presented. Results from the run employing the large dataset derived from the MDDR database (Dataset B) are given in Tables 11 and 12 and visualized in Figure 17. All results from this dataset employing Unity fingerprints are reproduced from Hert *et al.*³¹

First, we compared retrieval rates of Atom Environments and Unity fingerprints used in combination with the Tanimoto similarity coefficient. The percentage of active compounds found in the top 5% of the ranked database varies greatly between the datasets, from 95.04% in case of Renin inhibitors to 13.16% in the case of Cyclooxygenase inhibitors, if Atom Environment descriptors are employed. For Unity fingerprints, retrieval rates range between 80.54% in case of Renin inhibitors to 9.39% in case of Cyclooxygenase inhibitors. Retrieval rates averaged over all datasets using single queries gives better results for Atom Environments with 37.44% of the active compounds retrieved, compared to 30.58% if Unity fingerprints are used.

In the next series of runs, we used information from multiple molecules for similarity searching. Atom Environments, information-gain base feature selection and the Naïve

Bayesian Classifier were compared to results reported earlier using Unity fingerprints and data fusion as well as Binary Kernel Discrimination³¹. Results are again given in Table 11 and Table 12 and they are visualized in Figure 17.

Table 11. Mean percentage of active compounds found in the top 5% of the ranked library. Results using single Unity fingerprints, data fusion and Binary Kernel Discrimination are taken from Hert *et al.*³¹. Numbers in parentheses are standard deviations of the mean values.

Dataset	Unity, Single Queries	Atom Environ ments, Single Queries	Data- Fusion MAX	BKD 10 act., 10 inact., k=100	BKD 10 act., 100 inact., k=100	AE, 10 act., 10 inact., all features	AE 10 act., 100 inact., 250 Features
5HT3 Antagonists	21.15 (7.36)	25.40 (10.46)	49.03 (5.43)	47.79 (4.28)	52.32 (8.27)	61.67 (6.55)	66.58 (8.65)
5HT1A Agonists	18.43 (5.32)	27.73 (6.63)	37.15 (4.06)	30.78 (5.71)	38.19 (7.03)	44.19 (6.77)	57.05 (6.41)
5HT Reuptake Antagonists	24.02 (10.08)	22.75 (8.62)	49.68 (5.45)	37.28 (4.56)	45.82 (7.93)	41.40 (4.26)	46.07 (3.99)
D2 Antagonists	17.35 (6.60)	23.24 (11.09)	37.40 (4.92)	33.30 (7.70)	38.65 (7.38)	49.14 (6.94)	53.69 (7.52)
Renin Inhibitors	80.54 (13.83)	95.04 (2.83)	88.62 (1.90)	89.84 (5.95)	93.34 (1.35)	94.66 (0.86)	95.71 (0.68)
Angiotensin II AT1 Antagonists	48.04 (17.95)	68.01 (11.19)	80.44 (6.08)	82.19 (4.59)	84.47 (6.59)	93.62 (1.98)	95.05 (2.49)
Thrombin Inhibitors	33.51 (14.72)	34.79 (19.98)	58.58 (8.98)	54.48 (9.20)	63.06 (7.66)	60.14 (11.04)	66.15 (7.27)
Substance P Antagonists	26.87 (10.47)	31.03 (14.09)	47.14 (5.16)	44.79 (6.47)	58.39 (8.27)	59.16 (9.51)	68.43 (5.48)
HIV Protease Inhibitors	37.60 (13.82)	49.56 (25.04)	61.62 (7.85)	59.07 (9.73)	68.45 (8.31)	72.26 (11.41)	76.00 (4.60)
Cyclooxygenase Inhibitors	9.39 (4.76)	13.16 (6.49)	26.52 (7.15)	30.51 (6.58)	33.15 (4.68)	25.66 (4.85)	34.70 (4.47)
Protein Kinase C Inhibitors	19.42 (13.43)	21.13 (16.14)	48.01 (8.99)	47.47 (9.84)	49.37 (10.84)	50.50 (4.25)	54.61 (10.13)
Average	30.58 (10.76)	37.44 (12.05)	53.11 (6.00)	50.68 (6.78)	56.84 (7.12)	59.31 (6.22)	64.91 (5.61)

Table 12. Influence of the number of selected features on the mean percentage of active compounds found in the top 5% of the ranked library, applied to the atom environment similarity searching algorithm. Numbers in parentheses are standard deviations of the mean values.

Training Dataset	of	10 active compounds, 10 inactive compounds			10 active compounds, 100 inactive compounds		
		150	250	all	250	500	all
Number of Features Selected							
5HT3 Antagonists		60.49 (9.41)	59.65 (8.10)	61.67 (6.55)	66.58 (8.65)	59.72 (5.72)	45.93 (3.77)
5HT1A Agonists		42.73 (10.76)	43.44 (10.15)	44.19 (6.77)	57.05 (6.41)	49.84 (5.83)	32.20 (5.05)
5HT Reuptake Antagonists		35.64 (6.80)	35.50 (4.94)	41.40 (4.26)	46.07 (3.99)	42.09 (6.27)	26.50 (8.02)
D2 Antagonists		44.78 (11.85)	46.55 (9.90)	49.14 (6.94)	53.69 (7.52)	52.18 (5.45)	37.22 (5.32)
Renin Inhibitors		94.18 (0.83)	94.44 (0.78)	94.66 (0.86)	95.71 (0.68)	95.00 (0.67)	92.29 (1.04)
Angiotensin II AT1 Antagonists		91.51 (4.95)	92.45 (4.03)	93.62 (1.98)	95.05 (2.49)	94.39 (2.11)	88.23 (2.81)
Thrombin Inhibitors		55.40 (5.88)	50.84 (11.50)	60.14 (11.04)	66.15 (7.27)	62.62 (11.50)	55.86 (7.75)
Substance P Antagonists		64.34 (5.65)	49.47 (11.19)	59.16 (9.51)	68.43 (5.48)	62.10 (7.34)	56.12 (5.51)
HIV Protease Inhibitors		71.66 (8.88)	72.89 (9.56)	72.26 (11.41)	76.00 (4.60)	73.19 (4.39)	64.80 (8.05)
Cyclooxygenase Inhibitors		22.60 (8.41)	20.93 (5.95)	25.66 (4.85)	34.70 (4.47)	26.84 (7.48)	14.50 (4.29)
Protein Kinase C Inhibitors		46.65 (6.15)	46.13 (6.86)	50.50 (4.25)	54.61 (10.13)	51.11 (12.11)	44.03 (7.74)
Average		57.27 (7.23)	55.66 (7.54)	59.31 (6.22)	64.91 (5.61)	60.83 (5.81)	50.70 (5.40)

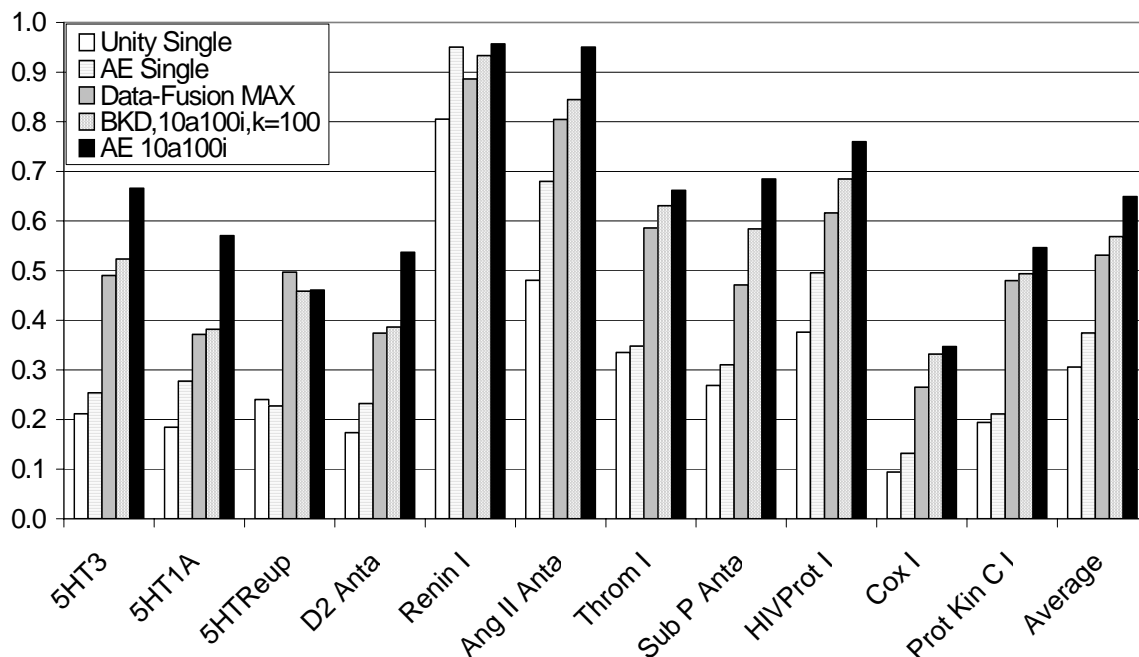


Figure 17. Fraction of active compounds retrieved in the top 5% of the sorted libraries. Compared are: Unity fingerprints (white) and Atom Environments (grey horizontal bars) using single query structures, both are used in combination with the Tanimoto coefficient. Data fusion (grey) and Binary Kernel Discrimination (BKD, dotted) are used to merge information from Unity fingerprints, and Atom Environments are for that purpose combined with the Naïve Bayesian Classifier (AE, black). Results using data fusion and Binary Kernel Discrimination are taken from Hert *et al.*³¹. Binary Kernel Discrimination and the Atom Environment method use 10 active and 100 inactive structures (AE 10a100i); the parameter k for the binary kernel method is set to $k=100$. Information-gain feature selection in case of the Bayesian Classifier was set to select 250 features.

Combining information outperforms results using single queries in all cases. The percentage of active compounds found in the top 5% of the ranked database again varies greatly between the datasets, from greater than 90% in case of Renin inhibitors and Angiotensin antagonists to about 30% in case of Cyclooxygenase inhibitors. Retrieval rates averaged over all datasets among the methods used gives lowest results for Binary Kernel Discrimination ($k = 100$) using 10 active and 10 inactive structures at 50.68% of the active compounds retrieved, followed by data fusion at 53.11%, Binary Kernel Discrimination using 10 active and 100 inactive structures ($k = 100$) at 56.84%, Atom Environments and the Naïve Bayesian Classifier using 10 active and 10 inactive structures at 59.31% and the same method using 10 active and 100 inactive structures at 64.91% actives found. Atom Environments and the Bayesian Classifier retrieve on average nearly 10% more active structures than the next best method, Binary Kernel Discrimination.

The highest number of active compounds is found in 10 out of 11 classes of active compounds by the Atom Environment approach if 10 active and 100 inactive compounds are used. Data fusion excels in one case, when applied to 5HT Reuptake inhibitors. If 10 active and 10 inactive structures are used, the Atom Environment approach comes first in 9 out of 11 classes of active compounds, with data fusion being superior in the case of 5HT Reuptake inhibitors. Binary Kernel Discrimination and data fusion are superior in the case of Cyclooxygenase inhibitors.

The influence of the number of selected features on similarity searching performance is given in Table 12. Feature selection in the case of 10 active and 10 inactive molecules gives approximately comparable results for any number of selected features, on average between 55% and 60% of the active compounds retrieved. If no feature selection is employed, 57.79% of all active structures are retrieved, so skipping feature selection in this case does not decrease search performance. If 10 active and 100 inactive structures are used for query generation, average retrieval rates vary between about 65% and 55%. If no feature selection is employed, the overall average retrieval rate falls even lower, to 50.70%. This result is clearly inferior to Unity fingerprints in combination with Binary Kernel Discrimination and shows the importance of using feature selection in combination with the Bayesian Classifier.

A consistent picture emerges in the validation using the large MDDR library (Dataset B), using single query structures with the Tanimoto coefficient. In 10 out of 11 cases (with the exception of 5HT reuptake inhibitors), Atom Environments outperform Unity fingerprints with respect to compound retrieval rates, with 37.44% vs. 30.58% of active compounds found. In absolute numbers, the performance difference is 6.86%. This calculates to a relative performance difference of 22.4%. Thus, Atom Environments capture more information that is relevant to the similarity searching task performed here.

If multiple query molecules are used, Atom Environments in combination with the Naïve Bayesian Classifier outperform Unity fingerprints in combination with data fusion as well as Binary Kernel Discrimination in most cases (9 or 10 out of 11 cases depending on the number of inactive compounds chosen, see results section). This observation is discussed in more detail below.

Comparing the performance of Atom Environments and the Naïve Bayesian Classifier to Unity fingerprints and Binary Kernel Discrimination, we find that in the case of 10 active and 10 inactive structures relative retrieval rates are 17.0% better if Atom Environments and the Naïve Bayesian Classifier are used. If 10 active and 100 inactive structures are

used, relative retrieval rates are on average 14.2% larger. In absolute numbers, Atom Environments and the Naïve Bayesian Classifier are superior by 8.63% (8.08%).

The absolute difference of retrieval rates between Atom Environments and Unity fingerprints is retained at about 7-8% if the Naïve Bayesian Classifier and Binary Kernel discrimination replace the Tanimoto coefficient, respectively. The relative performance gain drops from about 22% to 14-17%. This may be partly because retrieval performance gets close to the theoretical optimum in some of the datasets (Renin inhibitors and Angiotensin II inhibitors).

While the comparison of method until this stage compared different descriptors and machine learning methods, direct comparison is difficult precisely due to this reason, since performance differences cannot be easily deconvoluted into contributions of the descriptor and contributions of the particular machine learning method used. Therefore in the next step, the molecular representation of MOLPRINT 2D fingerprints was exchanged for ECFP_4 fingerprints and the classification was repeated, performing a ten-fold validation selecting 10 active and 10 inactive or 10 active and 100 inactive compounds, respectively. Results of this comparison are shown in Table 13. It can be observed that classification results are improved upon those employed before, on average from 59.31% retrieved compounds for MOLPRINT 2D descriptors to 66.67% of compounds for ECFP_4 fingerprints if 10 active and 10 inactive compounds are selected and all features are used for classification. If 10 active and 100 inactive compounds are used, selection of 500 features (corresponding to the number of features present in the active set) gives best results, retrieving 69.88% of compounds while before selection of 250 features was advantageous, identifying 64.91% of actives. It can be observed that the optimum number of selected features is shifted to a larger number for ECFP_4 fingerprints, which can possibly be explained by the larger degree of differentiation within the ECFP_4 atom types. Overall, performance could be improved by substituting MOLPRINT 2D descriptors for ECFP_4 fingerprints. The performance improvement amounted for the dataset containing 10 active and 100 inactive compounds to on average about 5% (4.97%).

Table 13. Influence of the number of selected features on the mean percentage of active compounds found in the top 5% of the ranked library, applied ECFP_4 fingerprints in combination with feature selection and the Bayes Classifier. Numbers in parentheses are standard deviations of the mean values.

Training Dataset	10 active compounds, 10 inactive compounds			10 active compounds, 100 inactive compounds		
Number of Features Selected	150	250	all	250	500	all
5HT3 Antagonists	63.61 (7.28)	66.31 (7.68)	70.75 (7.82)	66.44 (7.28)	74.39 (4.04)	57.82 (4.45)
5HT1A Agonists	51.77 (6.85)	54.35 (6.12)	57.16 (6.85)	59.73 (7.47)	62.42 (6.49)	45.17 (6.85)
5HT Reuptake Antagonists	45.56 (2.30)	42.98 (4.01)	49.00 (3.72)	48.71 (5.16)	52.72 (4.58)	42.69 (2.87)
D2 Antagonists	49.61 (2.86)	54.03 (3.12)	56.88 (5.45)	53.25 (6.75)	58.44 (4.42)	41.82 (6.75)
Renin Inhibitors	96.96 (0.54)	96.61 (0.45)	96.96 (0.80)	97.50 (0.63)	96.61 (0.63)	96.70 (0.63)
Angiotensin II AT1 Antagonists	97.21 (1.00)	96.46 (1.93)	97.53 (1.61)	97.86 (0.43)	97.43 (1.29)	95.50 (0.96)
Thrombin Inhibitors	74.40 (5.92)	78.06 (4.41)	74.65 (5.04)	83.35 (6.31)	80.33 (5.30)	70.62 (5.42)
Substance P Antagonists	58.33 (7.36)	57.85 (9.88)	66.42 (8.66)	66.42 (6.63)	72.33 (3.16)	55.58 (7.04)
HIV Protease Inhibitors	73.11 (3.65)	77.97 (3.51)	76.22 (3.64)	77.30 (2.97)	78.11 (3.92)	71.35 (3.92)
Cyclooxygenase Inhibitors	27.96 (5.11)	35.14 (6.39)	35.94 (4.63)	35.62 (4.79)	40.73 (7.99)	23.32 (3.99)
Protein Kinase C Inhibitors	46.61 (5.88)	49.32 (4.98)	51.81 (6.56)	55.20 (5.66)	55.20 (7.69)	43.21 (5.43)
Average	62.20 (4.42)	64.46 (4.77)	66.67 (4.98)	67.40 (4.92)	69.88 (4.50)	58.53 (4.39)

In order to establish now comparisons between different machine learning algorithms, in the following classification performance using ECFP_4 fingerprints in combination with different learning algorithms was evaluated. Performance data for Binary Kernel Discrimination and Data Fusion are taken from a recent publication²¹¹ and are only reproduced here. Results obtained using the Bayes Classifier are generated by the methods described in this work. Results are shown in Table 14 and visualized in Figure 18.

Table 14. Comparison of different machine learning algorithms for the identification of active compounds. In each case, structures are encoded employing ECFP_4 fingerprints and the percentage of compounds retrieved in the top 5% of the ranked compounds are reported.

Training Dataset	10 active compounds, 10 inactive compounds	10 active compounds, 100 inactive compounds	10 active compounds, 100 inactive compounds	10 active compounds, 100 inactive compounds
	Bayes	Bayes	BKD (from ref. ²¹¹)	Data Fusion (from ref. ²¹¹)
Number of Features Selected	all	500		
5HT3 Antagonists	70.75	74.39	65.3	72.2
5HT1A Agonists	57.16	62.42	58.7	64.2
5HT Reuptake Antagonists	49.00	52.72	50.3	49.7
D2 Antagonists	56.88	58.44	55.2	56.1
Renin Inhibitors	96.96	96.61	96.7	96.8
Angiotensin II AT1 Antagonists	97.53	97.43	98.0	97.4
Thrombin Inhibitors	74.65	80.33	74.9	74.7
Substance P Antagonists	66.42	72.33	67.3	62.2
HIV Protease Inhibitors	76.22	78.11	80.8	80.0
Cyclooxygenase Inhibitors	35.94	40.73	34.4	40.1
Protein Kinase C Inhibitors	51.81	55.20	49.6	57.8
Average	66.67	69.88	66.5	68.3

It can be observed that, on average, retrieval is very similar between the methods. BKD retrieves 66.5% of all active compounds, Data Fusion 68.3% while the Bayes Classifier (at a selection of 500 features) retrieves 69.88% of actives. From Figure 18 it can be seen that differences are much more profound between the different classes of active compounds than between the classification methods.

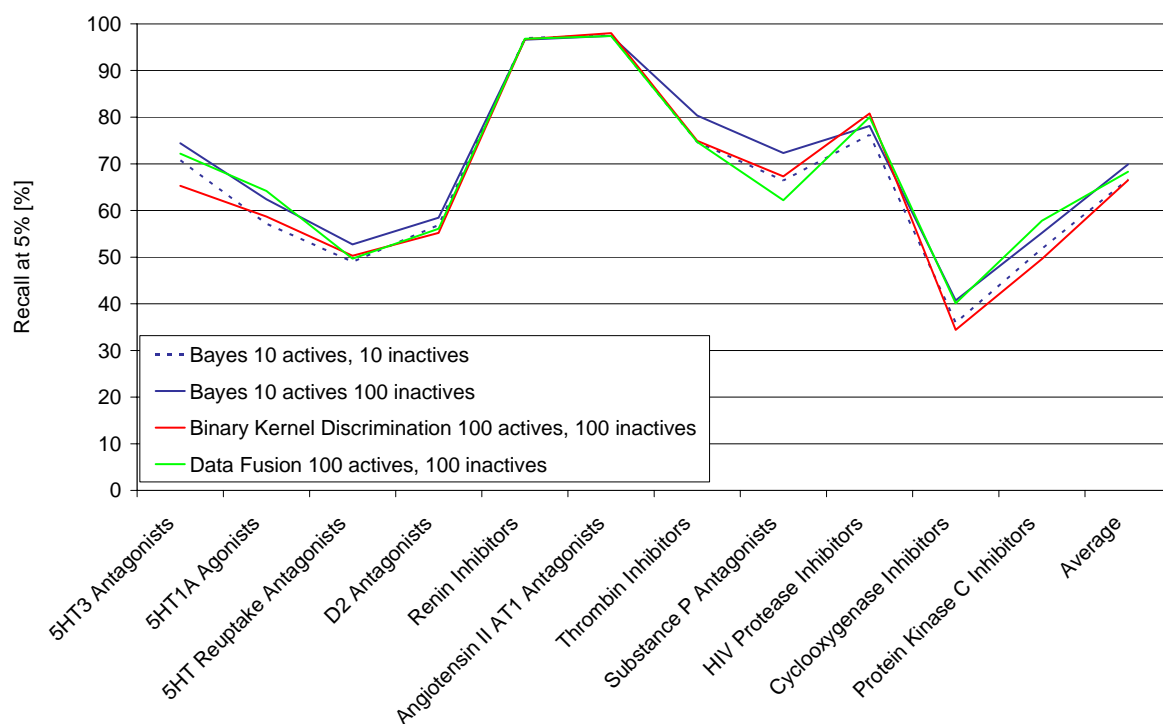


Figure 18. Percentage of active compounds retrieved employing ECFP₄ fingerprints and comparing the Bayes Classifier (selecting 10 actives and 10 or 100 inactives) to Binary Kernel Discrimination and Data Fusion. For the performance numbers selecting 10 active and 100 inactive compounds differences between the machine learning methods are minimal.

In addition to the influence of feature selection the question of how to deal with 0-probabilities in the context of the Naïve Bayesian Classifier needs to be answered. While a variety of approaches exist in the literature²⁴³ the question of how to apply the Laplacian correction in the context of molecular similarity searching needed to be explored in this work. Three approaches were compared, namely (a) ignoring features which are missing in one of the datasets, (b) applying a $1/m$ correction (where non-existent features are approximated by the inverse of the size of the dataset; this assumes that the feature is present in the dataset once) and (c) applying a $1/2D$ correction (where non-existent features are approximated by the inverse of the number of entries in the larger of the two datasets, multiplied by a factor of 2). In correction (c) the assumption is made that each feature that is not present would be present in a dataset of twice the size of the larger datasets of the two. Results are illustrated in Figure 19 and listed in Table 15.

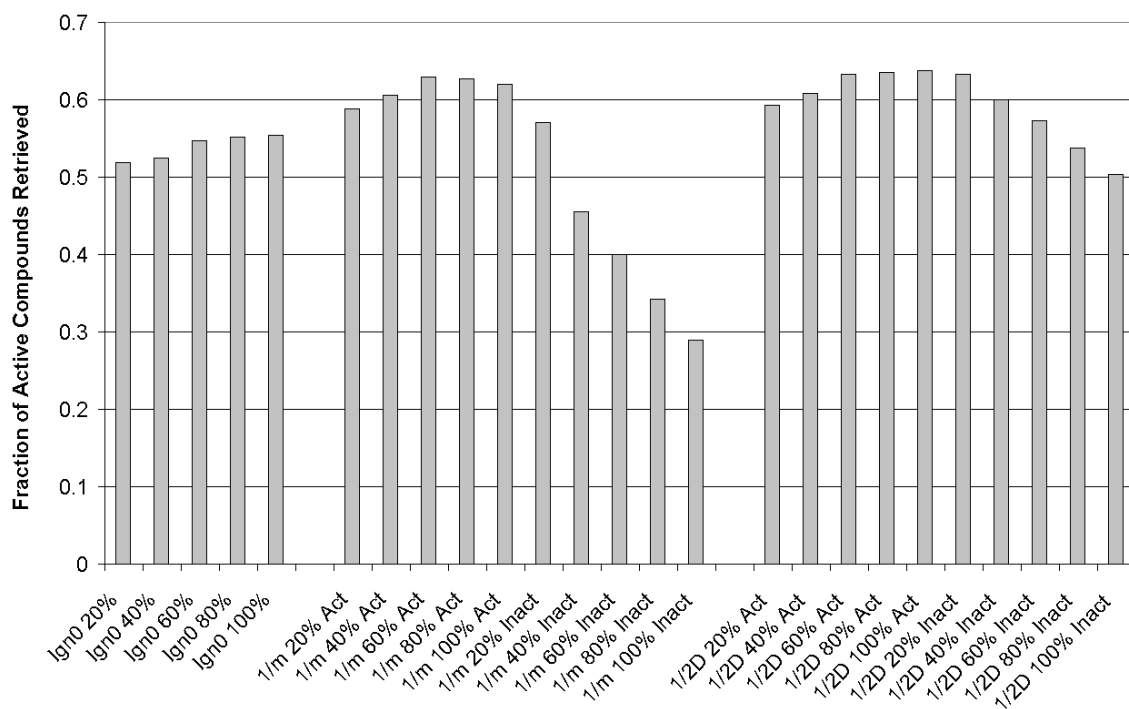


Figure 19. Performance of the classification method, depending on the Laplacian correction term employed. No correction (Ign0, omitting features present in only one of the datasets) shows worst performance, while the difference between the 1/m and the 1/2D corrections is the performance distribution, depending on the number of features selected. %Act and %Inact describe the percentage of features selected, relative to the total number of features present in either the active or the inactive dataset (which contain 10 and 100 structures, respectively).

Table 15. Performance of the classification method, depending on the Laplacian correction term employed. No correction (Ign0, omitting features present in only one of the datasets) shows worst performance, while the difference between the 1/m and the 1/2D corrections is the performance distribution, depending on the number of features selected. %Act and %Inact describe the percentage of features selected, relative to the total number of features present in either the active or the inactive dataset (which contain 10 and 100 structures, respectively).

Class	06233	06235	06245	07701	31420	31432	37110	42731	71523	78331	78374	Average
Ign0 20%	0.5850	0.5029	0.2338	0.4847	0.8817	0.8987	0.5544	0.4633	0.6364	0.0970	0.3731	0.5192
Ign0 40%	0.6022	0.5018	0.2533	0.4847	0.8984	0.9193	0.5439	0.4979	0.5986	0.0970	0.3729	0.5245
Ign0 60%	0.6212	0.5173	0.3438	0.4951	0.8984	0.9182	0.5397	0.5359	0.5873	0.1343	0.4258	0.5470
Ign0 80%	0.6388	0.5248	0.3562	0.4956	0.8880	0.9131	0.5257	0.5511	0.5503	0.1696	0.4532	0.5515
Ign0 100%	0.6439	0.5274	0.3559	0.4919	0.8776	0.9170	0.5086	0.5527	0.5455	0.1922	0.4771	0.5536
1/m 20% Act	0.6624	0.5536	0.2777	0.5104	0.9319	0.9192	0.6003	0.5765	0.7551	0.2299	0.4516	0.5880
1/m 40% Act	0.6722	0.5706	0.3266	0.5423	0.9346	0.9352	0.6038	0.6106	0.7359	0.2347	0.4937	0.6055
1/m 60% Act	0.6918	0.5965	0.4014	0.5416	0.9372	0.9413	0.6076	0.6523	0.7315	0.2863	0.5353	0.6293
1/m 80% Act	0.7089	0.5946	0.4106	0.5301	0.9326	0.9365	0.5873	0.6477	0.7031	0.2978	0.5486	0.6271
1/m 100% Act	0.7015	0.5875	0.4017	0.5226	0.9321	0.9309	0.5629	0.6447	0.6904	0.3054	0.5425	0.6202
1/m 20% Inact	0.6028	0.4823	0.3530	0.4706	0.9263	0.8974	0.5226	0.6328	0.6658	0.2347	0.4923	0.5710
1/m 40% Inact	0.4069	0.3077	0.2304	0.3218	0.9043	0.8268	0.4135	0.5074	0.6065	0.0810	0.4081	0.4558
1/m 60% Inact	0.3089	0.2357	0.1662	0.2683	0.8706	0.7749	0.3617	0.4316	0.5734	0.0425	0.3624	0.3997
1/m 80% Inact	0.2071	0.1524	0.1117	0.2122	0.8330	0.7012	0.3174	0.3520	0.5277	0.0244	0.3231	0.3420
1/m 100% Inact	0.1248	0.0848	0.0722	0.1579	0.7822	0.6279	0.2759	0.2945	0.4858	0.0145	0.2661	0.2897
1/2D 20% Act	0.6594	0.5569	0.2828	0.5260	0.9329	0.9196	0.6078	0.5851	0.7547	0.2535	0.4475	0.5933
1/2D 40% Act	0.6667	0.5781	0.3398	0.5436	0.9365	0.9333	0.6071	0.6129	0.7397	0.2351	0.5038	0.6088
1/2D 60% Act	0.6973	0.5934	0.4100	0.5403	0.9399	0.9377	0.6127	0.6675	0.7332	0.2909	0.5380	0.6328
1/2D 80% Act	0.7167	0.6000	0.4186	0.5296	0.9352	0.9362	0.5974	0.6667	0.7150	0.3252	0.5523	0.6357
1/2D 100% Act	0.7251	0.6001	0.4195	0.5379	0.9346	0.9365	0.5831	0.6751	0.7064	0.3427	0.5563	0.6379
1/2D 20% Inact	0.7163	0.5868	0.4097	0.5301	0.9328	0.9345	0.5692	0.6952	0.6988	0.3460	0.5484	0.6334
1/2D 40% Inact	0.6628	0.5455	0.3705	0.5018	0.9305	0.9152	0.5354	0.6612	0.6718	0.2874	0.5199	0.6002
1/2D 60% Inact	0.6170	0.4951	0.3387	0.4756	0.9211	0.9008	0.4997	0.6428	0.6576	0.2460	0.5068	0.5728
1/2D 80% Inact	0.5536	0.4282	0.3032	0.4257	0.9112	0.8864	0.4745	0.6008	0.6431	0.2035	0.4812	0.5374
1/2D 100% Inact	0.4945	0.3650	0.2665	0.3912	0.8935	0.8707	0.4567	0.5522	0.6254	0.1649	0.4543	0.5032

Ignoring features which are only present in one of the datasets shows worst performance, while the $1/m$ and the $1/2D$ corrections show no differences in the peak performance at the optimum number of selected features. Still, performance over the number of features selected shows great differences: While the selection of a number of features approximately equal to the number of features present in the smaller (the active) dataset is crucial for the $1/m$ correction, performance using the $1/2D$ correction shows a more even distribution. This is a result of the different dataset sizes: While very rare features assume the value $1/m$ for the smaller dataset, this value is often larger than the actual frequency encountered in the larger dataset, hence too high a probability is assigned here. In case of $1/2D$ correction a very small probability (smaller than any probability of the features present in any of the datasets) is assigned, showing increased performance of the methods due to proper ‘punishment’ of features rarely encountered.

Using the MAO data set (Dataset C), as shown in Figures 20 and 21, classification is comparable to Binary Kernel Discrimination in the case of a 50/50 split of training and test sets (Figure 20).

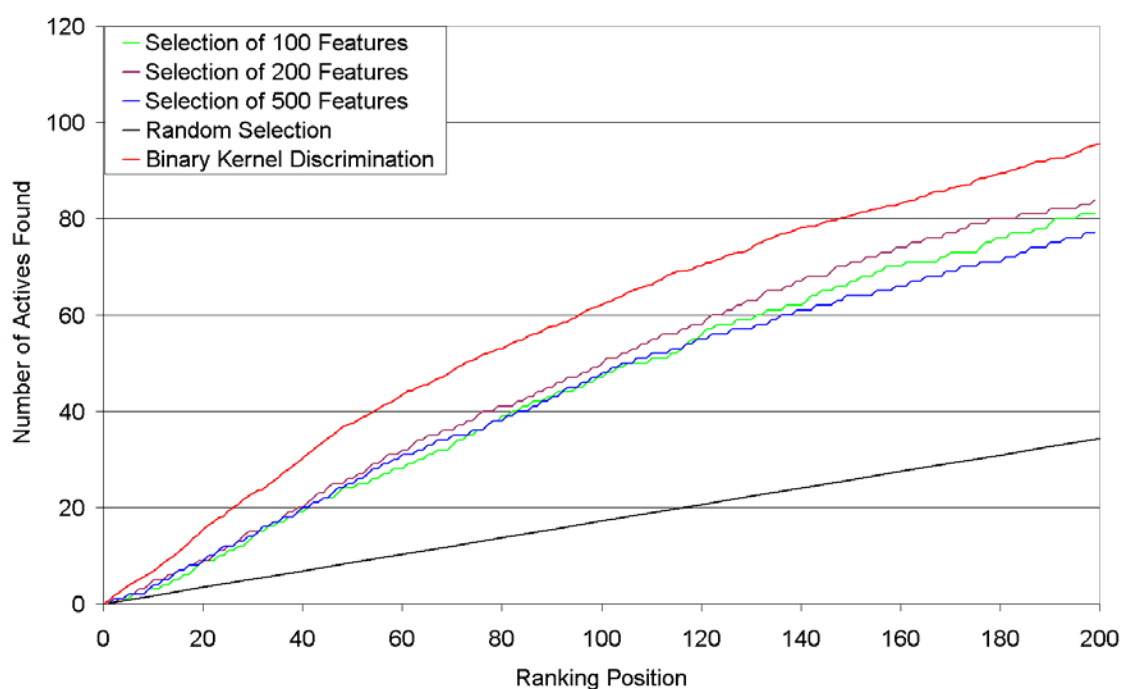


Figure 20. Number of MAO inhibitors compounds found at the top 200 positions of the ranked library when a 50/50 split of training and test data are used. The Bayesian Classifier and Binary Kernel Discrimination show comparable performance, with the Bayesian Classifier using 500 features showing slightly better performance from ranking position 40 to 160.

In the experiment with the smaller training set (Figure 21), comprising about 12% of the whole data set, Binary Kernel Discrimination is superior to the Bayesian Classifier. Where 50 active compounds are found at ranking position 100 in case of the Bayesian, the Binary Kernel Discriminator has already found 60 of the active compounds at this position.

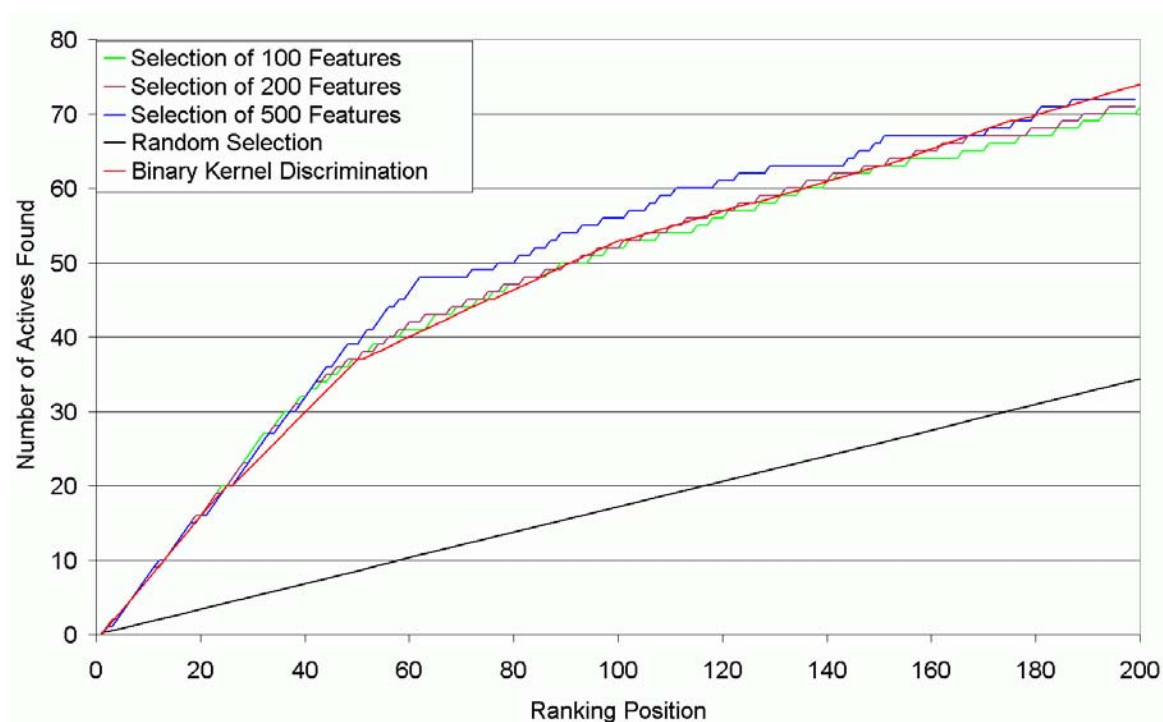


Figure 21. Number of MAO inhibitor compounds found at the top 200 positions of the ranked library when 200 training data points are used. Binary Kernel Discrimination outperforms the Naïve Bayesian Classifier on sparse training data when using this data set.

Applied to the MAO data set, Atom Environments perform comparably or in a slightly better way compared to Binary Kernel Discrimination where the (large) training set comprises 50% of the library. If a smaller number of 200 training data points are used, Binary Kernel Discrimination performs better than the Naïve Bayesian Classifier on this dataset. This is in contrast to the results obtained from the large MDDR dataset, where also small training data sets (10 active and 10 inactive structures or 10 active and 100 inactive structures) were used. Thus it may be attributed to the particular nature of the data set.

7.3 Conclusions

In this chapter 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 here can also be useful in identifying key functional groups in active molecules and is computationally efficient. There is ongoing research in substituting the Sybyl atom types by other descriptors (e.g. to include stereochemistry), employing fuzzy matching and other machine learning techniques that are expected to further improve the performance.

Applied to a recently published dataset using more than 100,000 molecules from the MDL Drug Data Report (MDDR) database, the Atom Environment approach consistently outperforms fusion of ranking scores (in 10 out of 11 cases) as well as Binary Kernel Discrimination in combination with Unity fingerprints (in 10 out of 11 cases if 10 active and 10 inactive structures are used for training and in all cases if 10 active and 100 inactive structures are available for training). Performance is particularly superior in the case of more diverse datasets, such as the 5HT3 set (which contains ligands of all receptor subtypes). Overall retrieval rates among the top 5% of the sorted library are nearly 10% better (more than 14% better in relative numbers) than the second best method, Binary Kernel Discrimination. Employing the same descriptors, ECFP_4 fingerprints, in every case, it is found that performance (selecting 10 active and 100 inactive compounds) is very similar when the Bayes Classifier is compared to Binary Kernel Discrimination and Data Fusion (where all methods retrieve between 66.5% and 69.9% of actives). Therefore, it might be a good idea to firstly pay attention to the amenability of the particular activity class to similarity searching, then look for the best descriptor available (which today seems to be circular ECFP_4 fingerprints) and finally choose a learning method which might show best performance for a particular class of activities, although the latter point seems to be marginal, compared to the first two.

The Laplacian correction was found to be important for the performance of the method presented here. Ignoring features which were present in only one of the datasets showed

consistently worse performance than either of the two Laplacian corrections employed. The nature of the Laplacian correction does not affect peak performance, but does considerably affect the variation in performance with the number of features selected. When used on a monoaminooxidase data set, the method performs as well as Binary Kernel Discrimination using atom pairs and topological torsion in the case of a 50/50 split of training and test compounds. In the case of sparse training data, Binary Kernel Discrimination is superior. This may be due to the particular data set, as it partly contradicts the finding on the first large and diverse MDDR dataset. Another possible explanation is that Binary Kernel Discrimination performs better with atom pairs and topological torsion than with Unity Fingerprints. Upon varying the number of selected features in the method presented, information-gain based feature selection is shown to be a crucial step for the performance of the Naïve Bayesian Classifier.

8 Surface-fingerprint based similarity searching

While the similarity searching approach based on circular fingerprints presented in the previous chapter shows significant retrieval of active structures, it at the same time retrieves a large number of analogous structures – structures which show a large degree of identical substructures. This shortcoming was addressed by us by employing a more abstract representation of the molecular structure, by representing a molecule by a set of local, but overlapping surface patches, each spanning about 8Å in diameter. While this approach still ignores long-distance information about the molecule, it at the same time (due to its local nature) alleviates one of the most common problems of three-dimensional descriptors, the problem of choosing the right conformation for the calculation of descriptors and, more generally, of the conformation-dependence of spatial descriptors.

Three-dimensional descriptors are generally created in a more complex and more computationally demanding process than two-dimensional descriptors. Compared to 2D descriptors, they have to deal with problems of translational and rotational variance as well as coping with the information overload resulting from a possible conformational explosion in 3D space. Still they possess the advantage of being able to identify molecules which exhibit similar properties (for example pharmacophores) in three-dimensional space without sharing 2D (connectivity table) similarity.

Descriptors that are invariant to both rotation and translation are known as TRI (Translationally and Rotationally Invariant) descriptors. Translational invariance can be achieved by using a coordinate system relative to the molecule and by centering the molecule with respect to it. Rotational invariance can be achieved by using distances between features instead of measuring coordinates in absolute space. This is the basis of autocorrelation approaches, which are well-known in both two dimensions^{108,244} and three dimensions^{245,246}.

“Surface point environments”, the descriptor introduced in this paper, are constructed in a three-step process (see Figure 22). Firstly, points on a “molecular surface” are computed. Secondly, interaction energies at surface points are calculated using hypothetical force field probes with varying parameters corresponding to different interaction types. Thirdly, interaction energies are encoded into descriptors, encoding only local information about interaction profiles in binary presence/absence features. In the experimental section, we will describe the descriptor in detail and also briefly summarize some of the descriptors which are most similar to surface point environments.

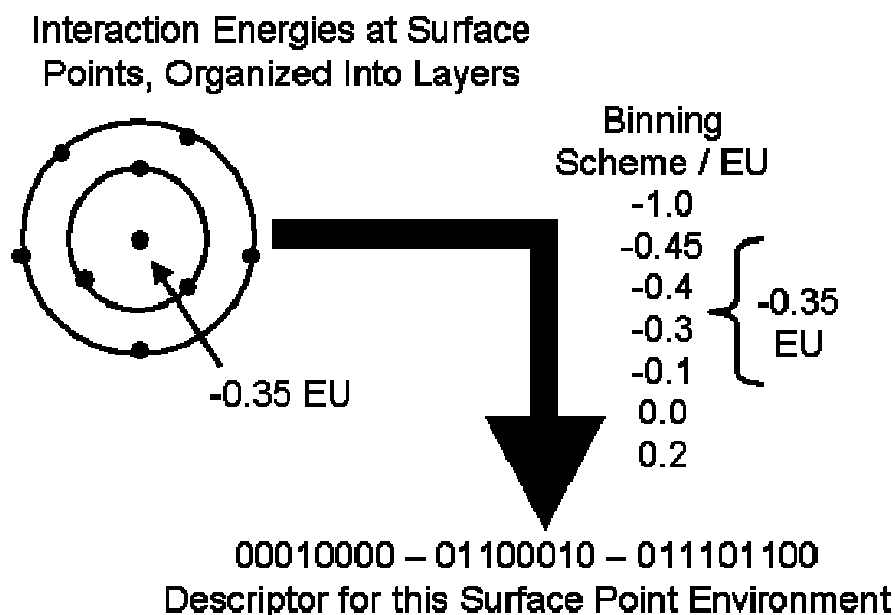


Figure 22. Illustration of the descriptor generation step. The surface point environment in the upper left is created for every point on the molecular surface. Interaction energies (given in EU, energy units) at every surface point are binned according to a binning scheme, which is calculated to give equally occurring bit frequencies in a random selection of molecules from the MDDR database and which is constant for every probe used. Bits are set in the final descriptor if interaction energies within a bin range are given in the particular layer. Hyphens separate parts of the descriptor created from different layers (distance ranges) of the surface point environment.

When descriptors are calculated for a molecule, its representation in (this) chemical space is defined. In particular in the world of 2D fingerprinting, feature vectors (fingerprints) are first calculated followed by similarity using agreements and discrepancies between all the computed features. Following the idea that most of the features calculated are (for our purposes) noise, a feature selection method is advisable. Here, we employ information-gain based feature selection as introduced by Quinlan²³⁰ as presented in the preceding chapter.

If information from more than one molecule is given, the problem of merging information can be difficult. Similarity coefficients have a shortcoming in that they, by nature, are only able to deal with single fingerprints. A simple method to combine information from multiple molecules, is to define minimum cutoff frequencies for a feature to enter the merged fingerprint³¹. Another recent approach is data fusion^{43,165}, which can be based on carrying out a series of single query similarity searches.

Here we follow a different route. In a fashion similar to Binary Kernel Discrimination^{207,233}, a type of data fusion is performed prior to scoring by using the Naïve Bayesian Classifier. This means that from all the information given in the representation step (and selected in the information-gain based feature selection step), a model is created that incorporates knowledge from all active structures. This contrasts with data fusion, as in data fusion single fingerprints are used for similarity searching and information from multiple searches is only fused after ranking. In contrast, by using the Naïve Bayesian Classifier, a unified model of active compounds is constructed prior to scoring.

To test MOLPRINT 3D we have utilized a data set containing 957 structures¹³² derived from the MDDR²⁴⁷ that contains active compounds from 5 different activity classes as described in the previous chapter. The set contains 49 5HT3 receptor antagonists, 40 Angiotensin Converting Enzyme (ACE) inhibitors, 111 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase inhibitors, 134 Platelet Activating Factor antagonists and 49 Thromboxane A2 antagonists. An additional 574 compounds were selected randomly from the MDDR database and did not belong to any of these activity classes. This dataset was previously examined¹³² using a variety of similarity searching methods and therefore serves as a suitable benchmark.

A number of tasks have been performed on this dataset. Firstly, parameters of the algorithm were optimized to achieve good performance in similarity searching. Performance in similarity searching tasks was then compared to other algorithms. As described above, translational and rotational variances as well as conformational tolerance are important points where 3D descriptors are used. The tolerance of this descriptor with respect to conformational variance was investigated using a sample from the MDDR data set of all groups of active compounds. Finally, selected features were projected back onto molecular space to investigate agreement with experimentally determined binding patterns. This was performed to check that results from similarity searching were not random patterns of surface points and in addition to investigate the use of this method for elucidating binding patterns in ligand-receptor complexes.

8.1 Materials and methods

a) Descriptor generation / molecular representation

The generation of surface point environments comprises four major steps which are summarized in Table 16. Firstly 3D coordinates are calculated from the two-dimensional

representation of the structure and saved in hydrogen-depleted SD format. This step is performed using Concord 4.0.7^{248,249} with standard settings.

Table 16. Main steps in descriptor generation, listing programs currently used in each step and exemplary important parameters. In principle, most parts of the algorithm are replaceable by a wide variety of programs.

Algorithm Step	Currently Used Program	Selected Important Parameters
Generation of 3D coordinates	Concord	
Calculation of Surface Points	msms	Sphere radius, probe size, triangulation density
Calculation of Interaction Energies	GRID	Probe (and various others)
Transformation of interaction energies into descriptors	Perl script	Binning, number of bins, threshold levels

The three-dimensional structure of the molecule is used as input for the triangulation of the molecular surface. The programme msms¹⁴⁵ was used to calculate the solvent-excluded surface with default radii multiplied by a factor of 2.0 to decrease sterical repulsion terms in the subsequent calculation of interaction energies. (The algorithm of msms is outlined in the following section.) This gives a representative (although not uniform) interaction surface with a spread of interaction energies evenly distributed across positive and negative values for all probes except the DRY probe (which due to its nature only gives negative interaction energies). Exemplary radii used are 3.08Å for nitrogen, 2.8Å for oxygen, 3.48Å for carbon and 2.4Å for hydrogen. The probe radius for the triangulation of the surface is set to 1.5Å, approximately corresponding to the radius of a water molecule. Triangulation densities are set to 0.5/Å² and 2.0/Å², giving about 400 and 2000 points for an average sized molecule, respectively.

It is known that the algorithm implemented by msms does not create equidistant points on the molecular surface³⁶. In order to achieve equidistant points on the surface, algorithms like GEPOL¹⁴⁶ may be employed instead. One should keep in mind though that there is no molecular microscopic equivalent of the macroscopic concept of a “surface”, so that parameter choices in this step are by and large arbitrary. In addition, liquid systems are

governed by dynamic changes of angles, distances and charges (by proton transfer) which underlines the fact that a molecular “surface” is only a crude approximation on a microscopic scale.

The SD file of the molecule is converted to hydrogen-added mol2 format using OpenBabel 1.100.2²³⁴. The mol2 file is converted to PDB format containing GRID atom types by the utility gmol2 that accompanies GRID^{35,103}. The three-dimensional coordinates calculated in the previous step are fed into the GRID input file grid.in employing the POSI directive in order to calculate interaction energies at the calculated surface points. The maximum energy (EMAX) is set to 5.0 in grid.in. The LEVL -1 directive is used to write GRID output in ASCII format, otherwise standard settings for GRID are used. Currently C3, DRY, N1+, N2, O and O- probes are used, which we expect to cover a variety of possible interactions between ligand and target. Definitions and characteristics of the probes are shown in Table 17 below.

Table 17. Probes employed to characterize the ligand surface, along with their definitions and the main interactions they are intended to “probe” for. While five of the probes are employed to detect areas with characteristic charge and hydrogen bonding properties, the DRY probe is probably the most atypical one in that it attempts to detect entropic contributions upon ligand binding, when water in the cavity is exchanged for the ligand.

Probe Symbol	Probe Definition	Main Interaction Probed For
C3	Methyl CH3 Group	Lipophilic Interactions
DRY	Hydrophobic Probe	Entropic Contributions When Water is Exchanged for the Ligand
N1+	sp ³ amine NH3 cation	Negative Charges
N2	Neutral flat NH2 (amide)	Hydrogen Bond Acceptor Functions
O	sp ² carbonyl oxygen	Hydrogen Bond Donor Functions
O-	sp ² phenolate oxygen	Positive Charges

The energy values calculated at the points on the molecular surface are binned using a Perl script. Binning of energy values is illustrated in Figure 22. For each point on the molecular surface, its topologically adjacent neighbours (as given by msms) are calculated

and arranged in layers. Points on the surface which are adjacent to the central point (“level 0”), for which the descriptor is generated in this particular step, belong to layer 1. Points which are adjacent to points in layer 1 belong to layer 2, excluding the central point. Points in layer n generally are those which are adjacent to points in layer $n-1$ and which have not been assigned to a layer of lower order.

In order to create binary presence/absence interaction energy ranges, equiproportional bin ranges have been calculated for the discretization of continuous interaction energies. A random selection of 53 structures from the MDDR database were chosen and surface points and interaction energies were determined with a triangulation density of $0.5/\text{\AA}^2$ and the C3, DRY, N1+, N2, O and O- probes. Cumulative frequencies of interaction energies were calculated. The seven bin thresholds were set to give equal populations to all eight bits. The resulting cut-off energies are given in Table 18. All bits corresponding to interaction energies present in a given layer are set in the bitstring.

Table 18. Energy cutoff values for binning of interaction energies at surface points into bits. Values are calculated to give equi-frequent bits in a random selection of molecules from the MDDR database.

Bin Cutoff	Probe Type					
	C3	DRY	N1+	N2	O	O-
Cutoff 1/EU	-1.45	-1.12	-4.30	-5.20	-1.90	-2.80
Cutoff 2/EU	-1.08	-0.72	-3.20	-3.70	-1.20	-2.10
Cutoff 3/EU	-0.85	-0.40	-2.30	-2.45	-0.95	-1.75
Cutoff 4/EU	-0.65	-0.08	-1.70	-1.80	-0.80	-1.45
Cutoff 5/EU	-0.50	-0.05	-1.30	-1.38	-0.65	-1.22
Cutoff 6/EU	-0.35	-0.01	-0.90	-1.00	-0.52	-0.90
Cutoff 7/EU	0.72	-0.001	-0.55	-0.75	-0.42	-0.60

Overall, for each point on the molecular surface a separate surface point environment vector is calculated. This vector encodes interaction energies at each point of the surface and its neighbouring points. Thus, it describes a local surface point environment which potentially facilitates (or reduces) ligand-target binding. No long-distance information is included in our descriptor, which paves the way for a conformationally tolerant description of the molecular surface. On the other hand it neglects information about

overall shape of the molecule. The whole molecule is described by a set of surface point environment vectors.

The surface point environment descriptor described here is the surface equivalent of the two-dimensional atom environment descriptor which has been described earlier^{78,79}. We will now briefly compare it to similar approaches.

Some of the best-known TRI descriptors are the GRIND³⁵ (GRid INdependent Descriptor) and the MaP (Mapping of atomic Properties) descriptor³⁶. Three-dimensional autocorrelation²⁴⁶ also shows resemblance to the method presented here.

The GRIND³⁵ descriptor is based on interaction energies of the molecule with a probe, which is positioned on a regularly spaced grid. Interaction energies are calculated on a continuous scale using the program GRID¹⁰³. The probes used are typically O and N1 for hydrogen bonding interactions and the DRY probe for lipophilic regions, but several dozen probes are pre-defined in GRID which cover a range of possible ligand-target interactions. All interaction energies at grid points are then clustered to simplify the descriptor. Distance ranges (“bins”) are defined and auto- and cross-correlations between interaction energies are calculated. Because only the maximum product of interaction energies enters the descriptor, back-projectability is achieved. In contrast to the GRIND descriptor, our approach explicitly uses points on the molecular surface and bins are replaced by neighbour / non-neighbour relationships between points in space. In the method presented here, encoding is stopped at a fixed number of layers of adjacent surface points, only covering about 3-8 Å in diameter (depending on the surface point density chosen). Interaction energies resulting from different probes are treated independently, in that a separate fingerprint for a given surface point is created for each individual probe used.

MaP³⁶ also uses points on the surface of the molecule. Employing a modification of the GEPOL algorithm¹⁴⁶, equally spaced points on the molecular surface are calculated and categorical putative interaction properties of the underlying atom type are assigned. Fuzzy counts are used to increment the bin corresponding to the given triplet of two properties and the distance between them as well as, to a lesser extent, neighbouring bins. The calculation of equally spaced surface points provides information about surface interaction properties as well as a size description. In contrast to the MaP descriptor, continuous variables from the GRID force field are employed (which are subsequently binned). Bins are replaced by neighbour / non-neighbour relationships between points in

space and only small parts of the molecule are encoded in each feature in the fingerprint. Fingerprints resulting from different probes are treated independently.

Surface Autocorrelation²⁴⁶ constructs a spatial autocorrelation vector on the molecular surface. The electrostatic potential is calculated and assigned to surface points, which is then encoded in a single surface autocorrelation vector for the whole molecule. In contrast, we construct individual descriptors for each point on the surface using different probes, each of which covers only part of the molecular surface.

b) MSMS Surface Point Generation

The MSMS algorithm is in this work employed to generate points on the molecular surface of the ligand. Since there is no exact equivalent of the macroscopic concept of a “surface” in the world of atoms and molecules the concept of a “surface” may be misleading. Nonetheless, it here describes a set of points which are located, in a way as uniformly as possible, in the interface region between ligand and protein target, in a region where different interactions can be characterized as contributing to the ligand-target affinity.

MSMS firstly constructs the reduced surface (described below), from which the solvent accessible and the solvent excluded surface are generated in subsequent steps.

1. Generation of the Reduced Surface

The name “Reduced Surface” might be a bit misleading since its surface concept deviates strongly from what is usually referred to as a surface of objects. Nonetheless, it contains all the information needed for the generation of the subsequent solvent excluded and solvent accessible surface and is therefore a very useful transformation step of the molecular coordinates. Three different parts contribute to the reduced surface. Firstly, the *faces of the reduced surface*, which are defined by a probe atom rolled over the molecular surface in positions where the probe touches (is locked by) three surrounding ligand atoms. In this case, the atom centres are called the *faces of the reduced surface*, the lines connecting them are the *edges* and the corresponding atom centres are the *vertices of the reduced surface*. The other two possibilities are that the probe atom touches two ligand atoms without colliding with a third or that it only touches a single ligand atom. In the first case, the resulting edge is called a *free reduced surface edge* (since the edge does not belong to any face) while in the second case (one atom touched

only) the resulting vertex is called a *free RS vertex*. The resulting polygons associated with all *fixed* positions of the probe, plus the *free edges* plus the *free vertices* define the reduced surface of a molecule which is sufficient for generating solvent accessible and solvent excluded surfaces.

2. Generation of the Solvent Accessible and Solvent Excluded Surfaces

After generation of the reduced surface, which is only an intermediary step, the actual solvent accessible and solvent excluded surfaces are constructed. Input for this second algorithms is a set M of n spheres along with their radii r .

In the init stage the initial reduced surface face to be treated is identified. It is either user-specified or defined algorithmically by the atom whose X-coordinate minus radius is smallest. Starting with the initial face, a probe is rolled over every pair of atoms until a third atom is hit. For each new face this step is repeated iteratively until all reduced surface edges have been treated. This identifies a construct called the “closed surface”.

a) Analytical Representation of the Solvent Excluded Surface (may be self intersecting)

In this step, three different surface types are assigned to each area of the closed surface, which are “spheric re-entrant” surfaces where three atoms are touched by the probe atom, “toric re-entrant” surfaces where two atoms are touched, and “contact faces” where only single atoms are touched by the probe. The first and latter of those surfaces have spherical shape and their treatment is further performed using template libraries, while the toric re-entrant surfaces need individual treatment. Also, radial singularities do need to be treated separately which are those cases where the probe cannot rotate around an edge of the reduced surface and gets “stuck” between two atoms. In this case additional treatment is performed, creating triangular approximations and the reader is referred to the original publication of Sanner and Olson¹⁴⁵ for details.

b) Removal of self-intersecting parts

Self-intersecting parts between a probe in a fixed position and the associated face of the reduced surface are singularities which need to be removed since the surface would otherwise be self-intersecting in those positions. Three different kinds of singularities can be distinguished here which all treated

separately by MSMS and for details on this step the reader is referred to the original article¹⁴⁵.

c) Triangulation

The final step of reducing discrete points on the molecular surface is the triangulation of the surface generated in the previous step. As previously described, three different types of surface areas exist, which are toric re-entrant faces, on one side limited by spheric re-entrant faces (where the probe touches three atoms) and on the other side limited by contact faces (the “free” atoms which are the only ones in touch with the probe). Firstly, toric re-entrant faces are triangulated algorithmically. Spheric re-entrant and contact faces are both defined by spheres and they are treated by using pre-triangulated template spheres with the same radius as the actual sphere representing the atom.

c) Feature Selection

Feature selection is only employed in combination with the Naïve Bayesian Classifier and multiple query structures. This step is skipped where the Tanimoto coefficient is employed.

The information content of individual surface point environments is calculated using the information gain measure of Quinlan^{230,250} and was described in detail in the previous chapter.

The information gain, I , is given by

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

where

$$S = - \sum_i p_i \log_2 p_i$$

S is the information entropy (which is defined analogously to entropy in real mixtures); S_v is the information entropy in data subset v ; $|D|$ is the total number of data sets; $|D_v|$ is the number of data sets in subset v and p is the probability that a randomly selected molecule of the whole data set (or subset in case of D_v) belongs to each of the defined classes.

d) Classification

Two methods were employed for classification, the conventional Tanimoto coefficient and the Naïve Bayesian Classifier.

The Tanimoto coefficient⁴ is a symmetrical similarity coefficient, which takes both similar and dissimilar properties of two items to be compared into account. In case of binary feature vectors (which are given here, surface point environments are either present or absent in each molecule) the Tanimoto coefficient T_C can be written as

$$T_C = \frac{AND}{OR}$$

where AND is the number of features which are present in both feature vectors to be compared and OR is the number of features which are present in only one of the feature vectors. Features which are present in none of the vectors are neglected by this coefficient.

On the other hand, a Naïve Bayesian Classifier²³¹ was employed as a classification tool (which was described in detail in the previous chapter). In short, a Bayesian Classifier predicts the class that a new feature vector belongs to as the one with the highest probability of $P(CL_v | F)$ which is given by

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

where

$P(CL_v)$: probability of class v

$P(F)$: feature vector probability and

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

v : class.

For two datasets, after applying the assumption of independence of features, the resulting binary Naïve Bayesian Classifier is given by

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

This equation is used to perform relative scoring 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 (e.g. the class of active molecules). Molecules with the lowest values are most likely to belong to class 2 (e.g. the class of inactive molecules). The prior, $P(CL_1)$ and $P(CL_2)$ in formula 2, is set to the relative training set sizes.

e) Dataset Preprocessing

Salts and solvent were removed, if present. Structures were converted to SD format using OpenBabel²³⁴ 1.100.2 with the `-h` option to add all hydrogen atoms. Only the neutral forms of molecules were considered. Surface fingerprints were calculated directly from SD files. The 49 structures of the 5HT3 dataset and the 40 structures from the ACE dataset were converted correctly. From the PAF dataset, 2 out of the original 134 structures were not converted, leaving 132 structures. 1 of the 49 structures from the TXA2 dataset and 14 out of 574 structures from the “inactive” dataset were not converted, leaving 48 and 560 structures, respectively. Overall, descriptors for 937 of 957 structures were calculated. Failure was in all cases due to msms which produced core dumps. Replacement by a different algorithm might reduce the failure rate.

8.2 Results and discussion

To better understand behaviour of the surface point generation step the coordinates of surface points generated by msms¹⁴⁵ were analyzed: For point densities of $0.5/\text{\AA}^2$, $1.0/\text{\AA}^2$ and $2.0/\text{\AA}^2$, distances to all nearest neighbours of each individual surface point were calculated and density functions of nearest neighbour distances were plotted for corticosterone (Figure 23 and Table 19). The mean distance between points decreases from 1.74 to 1.37 to 1.09 Å if surface point densities are increased from $0.5/\text{\AA}^2$ to $1.0/\text{\AA}^2$ and $2.0/\text{\AA}^2$. Median distances decrease from 1.57 to 1.12 to 0.79 Å in this case. Point densities for other compounds show comparable distributions.

Table 19. Mean and median distances and first and third quartiles of inter-point distances between points on the molecular surface.

Point Distance in Å	$0.5/\text{\AA}^2$	$1.0/\text{\AA}^2$	$2.0/\text{\AA}^2$
Mean	1.74	1.37	1.09
1. Quartile	1.17	1.02	0.73
Median	1.57	1.12	0.79
3. Quartile	1.88	1.33	0.95

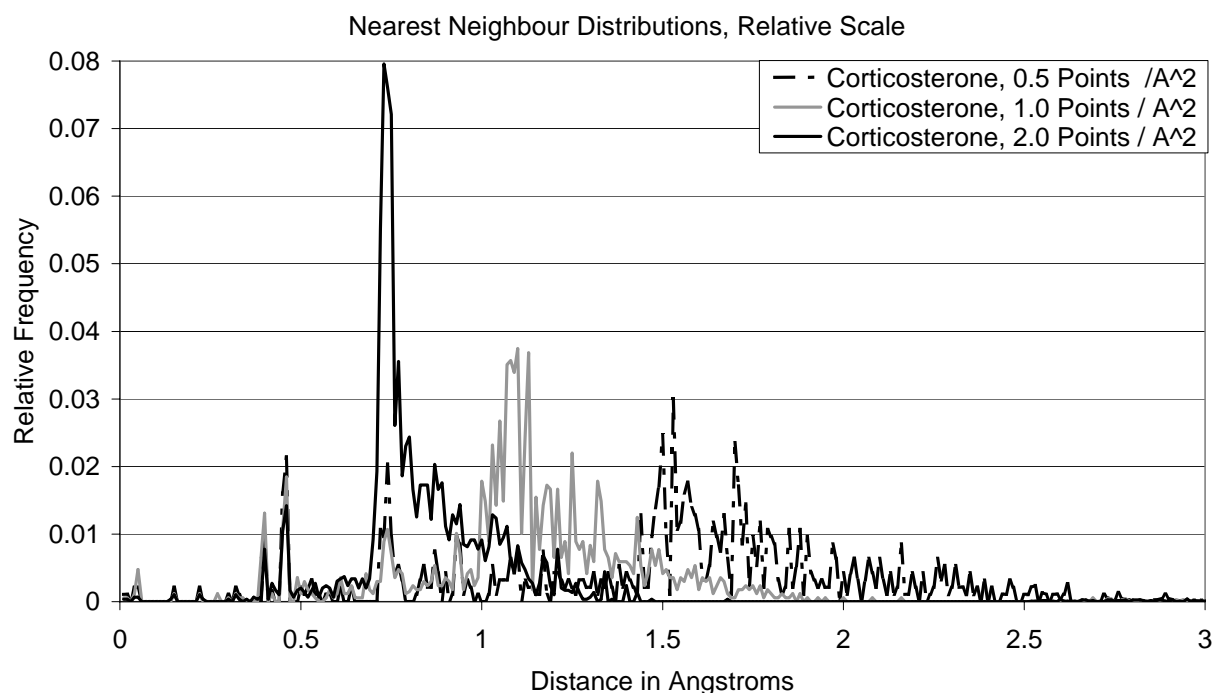


Figure 23. Distribution of surface points generated by msms. Displayed are distributions of nearest-neighbour distances on the surface of Corticosterone at triangulation densities of 0.5, 1.0 and 2.0 points / Å². The higher the point density chosen, the smaller the nearest-neighbour distances become. In each case, considerable spread of distances is observed. Distributions show comparable means and distributions for other molecules as well.

The distribution of points on the molecular surface is not equidistant but shows considerable spread, in particular in the case of smaller point densities. The distributions are remarkably similar among different individual compounds (data not shown). The average distance between points of around 1Å at point densities of 2.0/Å² amounts to an area of the molecular surface covered by each individual descriptor (layer 0-4) which spans about 8 Å in diameter.

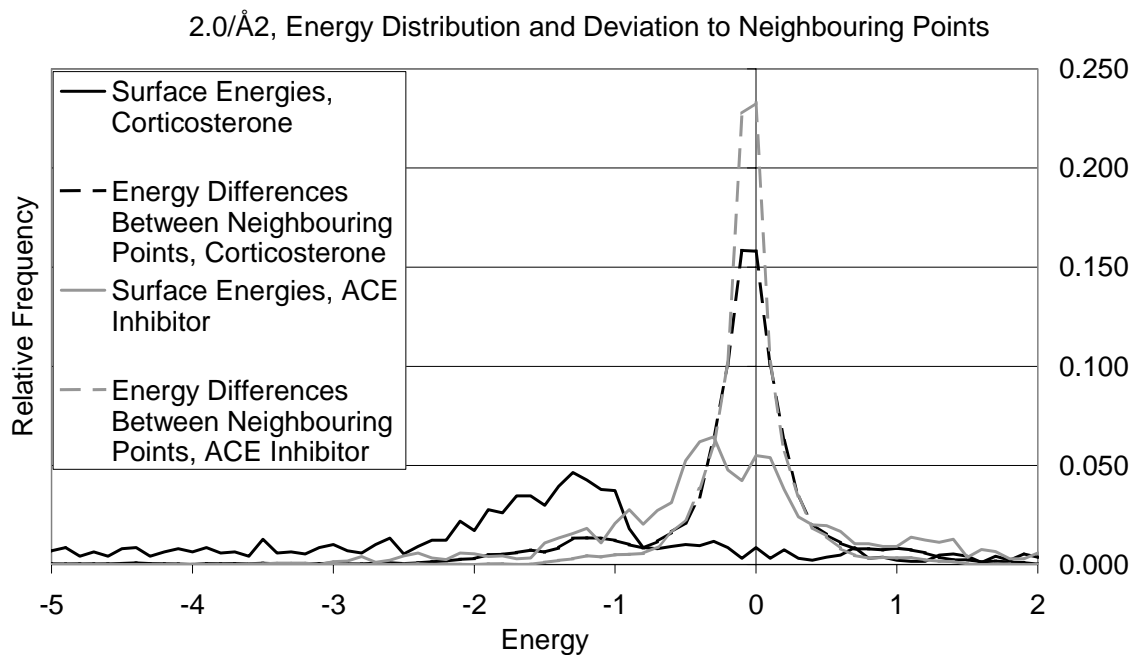


Figure 24. Energy distribution for corticosterone and an ACE inhibitor (both structures are given in Figure 25) using 2.0 points / Å² and the O- probe. Distributions of nitrogen probes shift the ACE distribution more to the left (compared to the ACE inhibitor). Dashed lines show energy differences between nearest neighbours and show smooth energy transitions from a point to its nearest neighbour.

For all six probes used by GRID, the energy distribution (relative frequencies of surface points within a certain energy range) over the molecular surface was calculated for a rigid molecule, corticosterone, and a more flexible ACE inhibitor at 2.0/ Å² grid spacing and an O- probe (Figure 24; see Figure 25 for structures). In addition, average energy changes from each point to its neighbours (smoothness of the potential) were calculated.

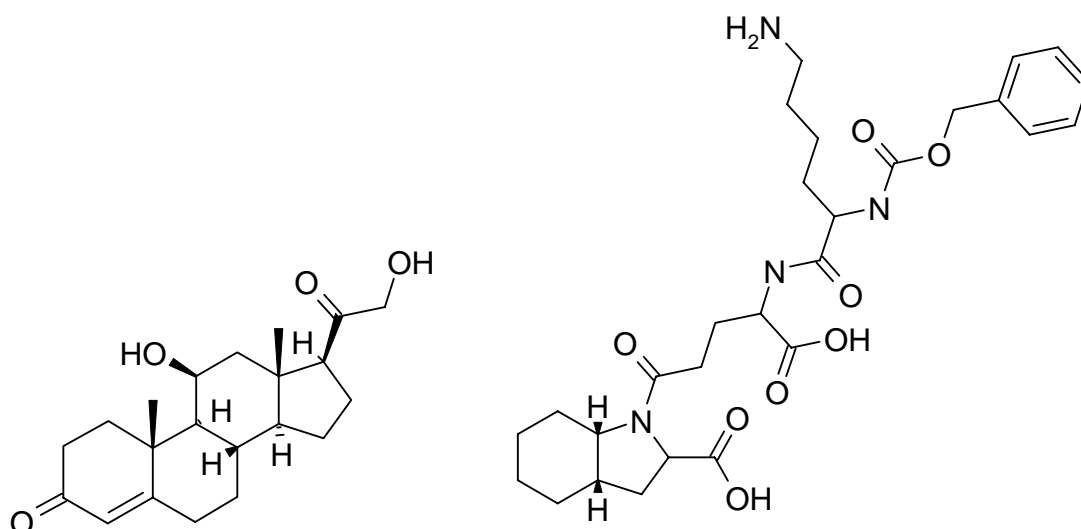


Figure 25. Corticosterone and the ACE inhibitor for which energy distributions using point spacing of $2.0/\text{\AA}^2$ and an O- probe are shown in Figure 24.

Absolute energy distributions show a maximum at a higher value for the ACE inhibitor than for corticosterone. Since the O- probe used is negatively charged and the maximum is shifted to higher (more unfavourable) energies for the ACE inhibitor, this is consistent with the larger number of negatively polarized oxygen atoms in this structure, relative to the total molecular surface area. Energy differences between individual points and their next neighbours show a sharp peak around the origin of the coordinate system, indicating that most changes in potential between points occur gradually. This shows that the energy functions are behaving smoothly which decreases the probability of artifacts in the descriptor generation step. Distributions for nitrogen (N1+ and N2) probes shift energy distributions for both compounds in opposite directions. The C3 probe gives approximately overlapping distributions for both compounds. Different probes show, as expected, different energy distributions.

In the first series of calculations, a ten-fold random selection of single structures from each dataset of active compounds was performed. The average hit rate for each class of active molecules (= the number of molecules among the 10 most similar structures belonging to the same activity class as the query structure) was calculated for each compound. Performance, defined as average hit rates, was compared for surface point environments calculated with a triangulation density of $0.5/\text{\AA}^2$ and $2.0/\text{\AA}^2$. In the second calculation, the Tanimoto coefficient was replaced by information-gain based feature selection and the Naïve Bayesian Classifier for classification. Ten-fold random sets of 5

active molecules have been selected and the set of inactive molecules was used in a 50/50 split⁷⁹. Feature selection was set to select 200, 500 or 1000 features for each set of molecules. As in the previous calculation, the hit rate among the ten highest ranked hits of the sorted library was calculated. Hit rates for the surface point environment descriptor and atom environments in combination with the Tanimoto coefficient are given in Table 20.

Overall hit rates are best for the 2D descriptor (atom environments), which on average retrieves 7.5 active compounds among the 10 structures most similar to the query (bottom of Table 20). Surface point environments created with a point density of $2.0/\text{\AA}^2$ are second in performance with on average 6.2 structures retrieved (if layer 0 – 4 are used for descriptor generation). At a lower point density of $0.5/\text{\AA}^2$ on average 6.1 structures are retrieved (if layers 0 – 1 are used for descriptor generation). If a point density of $0.5/\text{\AA}^2$ is employed, performance is broadly constant at 5.68 – 6.08 if the number of layers used to generate the descriptor is varied. This is true with the exception of employing single surface points; in this case, performance drops rapidly to an average hit rate of 1.88. If a point density of $2.0/\text{\AA}^2$ is employed, using layers 0 – 4 for descriptor generation gives best results, with an average hit rate of 6.2. Performance is again broadly constant at 5.32 – 6.24 if the number of layers used to generate the descriptor is varied. This is also true with the exception of employing single surface points; in this case, performance drops rapidly to an average hit rate of 2.22.

Table 20. Comparison of performance of the atom environment descriptor and the surface point environment descriptor in combination with all probes, both used in combination with the Tanimoto coefficient. Given are mean hit rates among the ten most similar compounds of a random selection of ten active compounds of each active dataset (standard deviation in parentheses). In the case of surface point environments, the number of layers used for descriptor generation and point densities of $0.5/\text{\AA}^2$ and $2.0/\text{\AA}^2$ is varied.

Layers Used	Point Density	5HT3	ACE	HMG	PAF	TXA2	Average
0	0.5/Å ²	1.80	0.90	3.0	2.10	1.60	1.88
		(1.40)	(0.88)	(2.16)	(1.37)	(1.46)	(1.46)
	2.0/Å ²	1.70	1.80	2.80	3.30	1.50	2.22
		(0.67)	(2.00)	(1.69)	(1.25)	(1.65)	(1.45)
0-1	0.5/Å ²	4.90	4.80	6.30	7.50	6.90	6.08
		(3.03)	(2.20)	(2.36)	(2.55)	(1.91)	(2.41)
	2.0/Å ²	4.30	2.60	6.90	7.10	5.70	5.32
		(2.45)	(1.96)	(2.77)	(3.03)	(2.26)	(2.49)
0-2	0.5/Å ²	5.50	5.80	3.60	7.60	7.00	5.90
		(2.37)	(2.82)	(2.84)	(3.24)	(2.16)	(2.68)
	2.0/Å ²	4.30	3.90	5.20	7.60	7.50	5.70
		(2.67)	(2.18)	(3.08)	(2.59)	(2.17)	(2.54)
0-3	0.5/Å ²	5.60	6.40	4.00	7.60	6.50	6.02
		(2.22)	(2.95)	(3.06)	(1.96)	(1.96)	(2.72)
	2.0/Å ²	5.70	5.00	4.70	7.70	7.50	6.12
		(2.16)	(2.62)	(2.45)	(2.79)	(2.22)	(2.45)
0-4	0.5/Å ²	5.40	5.40	4.00	7.30	6.30	5.68
		(2.22)	(2.55)	(3.02)	(3.27)	(2.21)	(2.65)
	2.0/Å ²	6.10	6.10	4.30	7.60	7.10	6.24
		(1.85)	(2.42)	(2.54)	(3.03)	(2.47)	(2.47)
Atom Environments		7.4	7.8	8.6	7.7	6.6	7.5 (2.3)
		(2.2)	(2.6)	(2.1)	(2.3)	(2.2)	

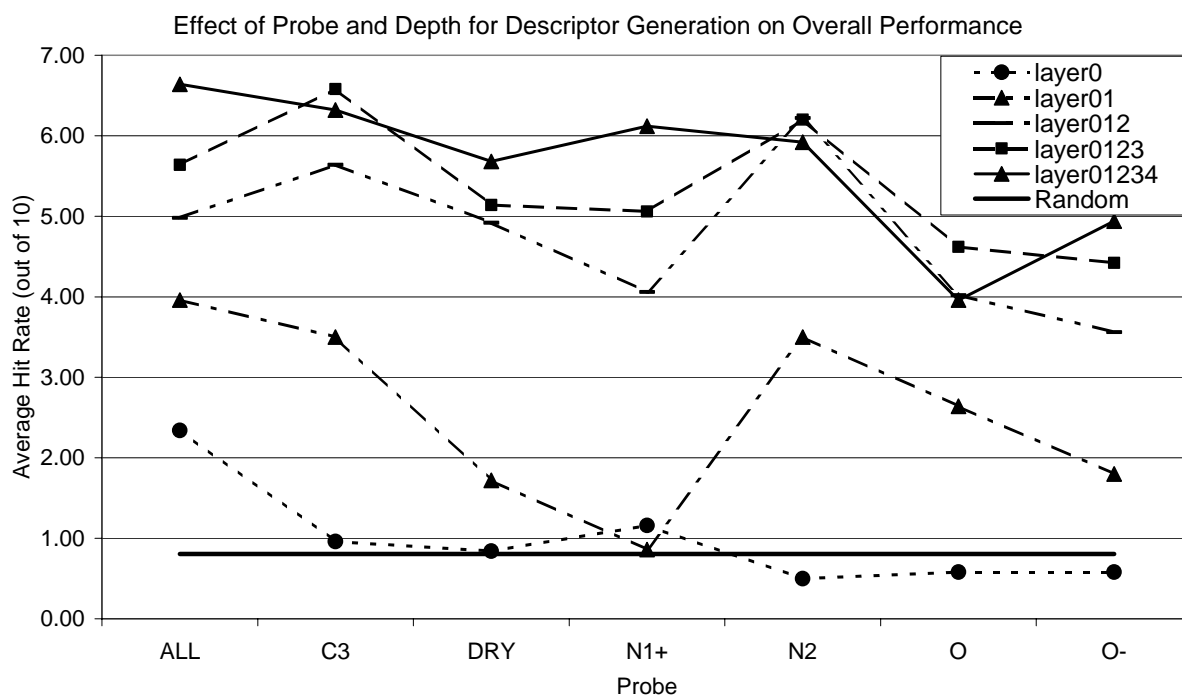


Figure 26. Similarity searching performance upon varying layer depth and choice of probe for descriptor generation. Performance is given using 500 features in each case and employing the Bayesian Classifier for classification. Performance depends on the choice of probe mainly in those cases of a small number of layers and levels off between three and four layers.

If single points are used (“level 0”), only a slight increase of performance over random selection can be observed (Table 20 and Figure 26). This means that the number of surface points with a given interaction energy with the probe (a property roughly analogous to measures like polar surface area, measuring the fraction of the surface with a given property) is not sufficient to achieve classification. Finer point spacing may be better at capturing local properties, although differences are minimal. Overall, significant enrichment is observed for each of the point densities chosen above (except for single points). Performance increases until all points up to layer 4 are incorporated into the descriptor and levels off at that point. This behaviour can be explained by the way surface point environment descriptors are generated. Bits in the feature vector represent presence/absence of interaction energies in a given energy range in a given layer. The further out the layer from the central surface point, the more points are present in that layer. In layer 4, a high number of (typically about 40 to 50) surface points are present. More often than not, all interaction energies are present in that layer, setting all bits in the vector to “1”. Therefore, no new information enters the descriptor in this layer. This is

even less likely in layers of higher order (as they contain an even larger number of points).

The superiority of the two-dimensional descriptor compared to its three-dimensional analogue (Table 20) purely with respect to hit rates is in agreement with earlier findings⁸⁴. In this earlier work descriptors of different dimensionality (2D and 3D) were compared, in combination with different clustering methods, with respect to their ability to cluster compounds showing the same bioactivity together. It was found that 2D descriptors (in combination with hierarchical clustering) are best at separating actives from inactives, given a particular target. Structural keys, hashed fingerprints and different 3D descriptors (including Unity 3D descriptors and putative 3-Point-Pharmacophores) were compared and all of MACCS keys, Daylight and Unity fingerprints were better at separating actives from inactives than any of the 3D descriptors used. In a more recent study²⁵¹ both 2D (Daylight) and 3D (3-Point-Pharmacophore) descriptors were found to show encouraging retrieval rates, with the 3D counterparts being able to retrieve actives also in case of low 2D similarity (where Daylight fingerprints failed). Still, in relative numbers, surface point environments retrieve only 17% and 24% fewer active compounds than atom environments, which still compares favourably with a number of 2D methods (Figure 26). Dependence of retrieval performance upon varying parameters of this algorithm is given in Table 21. Results obtained if all field fingerprints are used and a fixed number of 500 features is selected are visualized in Figure 27. Performance is given using the Bayesian Classifier and sets of 5 active molecules and a 50/50 split of inactive molecules.

Table 21. Similarity searching performance upon varying the layer depth used for descriptor generation, choice of probe for generation of interaction energies and number of features selected. The Bayesian Classifier is used for classification. Performance increases continuously from using only layer 0 and levels off if central points up to points 4 layers apart are used for descriptor generation. Performance is given for a point density of $0.5/\text{\AA}^2$. Values in parentheses are standard deviations from the mean values.

Layers Used	Number of Features	<i>Point Density 0.5/Å²</i>						
		Interaction Probe						
		ALL	C3	DRY	N1+	N2	O	O-
0	all	2.34	0.96	0.84	1.16	0.50	0.58	0.58
		(0.38)	(0.52)	(0.34)	(0.72)	(0.32)	(0.86)	(0.74)
0-1	200	4.02	4.06	2.90	2.46	4.32	2.44	2.36
		(1.92)	(1.54)	(1.78)	(1.42)	(1.60)	(1.22)	(1.28)
	500	3.96	3.50	1.72	0.86	3.50	2.64	1.80
		(1.52)	(1.84)	(1.06)	(0.58)	(1.84)	(1.02)	(0.50)
0-2	1000	4.00	3.18	1.72	0.70	2.18	1.96	1.70
		(1.62)	(1.46)	(1.20)	(0.38)	(1.44)	(0.94)	(0.52)
	200	4.96	5.14	4.80	3.60	5.72	5.16	3.40
		(2.00)	(1.36)	(1.68)	(1.68)	(1.64)	(2.06)	(1.92)
0-3	500	4.98	5.64	4.92	4.06	6.22	4.02	3.56
		(1.50)	(1.68)	(1.68)	(1.16)	(1.58)	(1.78)	(1.44)
	1000	5.38	6.00	4.08	3.16	5.90	3.82	2.74
		(1.24)	(1.48)	(1.68)	(1.86)	(1.90)	(1.38)	(1.34)
0-4	200	5.86	6.12	5.38	5.10	5.66	4.80	4.34
		(2.14)	(2.02)	(1.52)	(1.88)	(1.46)	(1.86)	(2.40)
	500	5.64	6.58	5.14	5.06	6.20	4.62	4.42
		(2.44)	(1.86)	(1.68)	(1.76)	(1.40)	(2.02)	(1.60)
0-4	1000	4.78	6.66	5.02	4.62	6.48	4.26	3.76
		(2.06)	(1.34)	(1.82)	(1.52)	(1.22)	(1.56)	(1.84)
	200	7.16	5.64	4.92	5.38	4.68	4.10	4.24
		(1.64)	(2.22)	(1.88)	(1.98)	(1.62)	(1.46)	(1.76)
0-4	500	6.64	6.32	5.68	6.12	5.92	3.96	4.94
		(2.00)	(1.86)	(1.86)	(1.68)	(1.76)	(1.88)	(2.16)
	1000	6.50	6.76	4.82	5.16	6.24	4.58	4.66
		(2.28)	(1.58)	(1.74)	(1.84)	(1.90)	(1.66)	(1.98)

Combining information from all interaction fields used (“ALL” column in Table 21) improves results over those obtained using only single interaction fields (other columns with probe names) in case of the best overall retrieval rate, using layer 0-4 and 200 features. Still, performance using only the C3, DRY, N1+ and N2 probes is surprisingly good and, depending on the precise parameters, often comparable to the performance

achieved with all probes. A possible explanation is that every probe simply describes the same variance in the data. Although different interaction energies are assigned to the same point in space if different probes are used, the overall variance (which is essential for classification) remains similar. A positive and a negative charge may give the same information, simply with an opposite sign.

Feature selection does not influence results at a small number of layers used for descriptor generation, but improves results throughout if more than layer 1 is used for descriptor generation. Feature selection continuously improves classification results if more than layers 0 and 1 are employed for descriptor generation (Table 5) and this is in analogy to atom environments if a large number of feature vectors is employed⁸⁰.

The difference in performance between atom environments and surface point environments depends on the class of active compounds. Performance on the 5HT3, ACE and PAF datasets is better for the 2D atom environments. Both 5HT3 and ACE datasets give on average 6.1 ($2.0/\text{\AA}^2$) and 5.4 ($2.0/\text{\AA}^2$) hits for surface point environments, compared to 7.4 hits (5HT3) and 7.8 hits (ACE) for atom environments. On the PAF dataset, surface point environments retrieve on average 7.6 ($2.0/\text{\AA}^2$) and 7.3 ($0.5/\text{\AA}^2$) hits vs. 7.7 in case of atom environments. For the TXA2 dataset results are comparable; the hit rates are 7.1 ($2.0/\text{\AA}^2$) and 6.9 ($0.5/\text{\AA}^2$), where atom environments retrieve on average 6.6 hits. Atom environments retrieve twice as many active compounds from the HMG dataset though; hit rates are 8.6 for atom environments vs. 4.3 and 4.0 for surface point environments at high and low point density, respectively. Surprisingly, performance of surface point environments on the HMG dataset shows much better performance if only layers 0-1, corresponding to much smaller surface patches, are used for classification. Elimination of 2D similar molecules from the HMG dataset did not give the expected result that high connectivity similarity favours the 2D method on this dataset in particular. The underlying reason for different performance of surface point environments and atom environments on this dataset is as yet unknown.

Table 22. Performance of the similarity searching algorithm using surface fingerprints in combination with the Tanimoto coefficient and the Bayes Classifier. The number of layers, the point density and (in case of the Bayes Classifier) the number of features selected are varied. Values are averaged over all datasets and numbers in parentheses are standard deviations from the mean values.

Method	Point Density	Layers Used	0	0-1				0-2			0-3			0-4		
Tanimoto	0.5/Å ²		1.88 (1.46)	6.08 (2.41)				5.90 (2.68)			6.02 (2.71)			5.68 (2.65)		
	2.0/Å ²		2.22 (1.45)	5.32 (2.49)				5.70 (2.54)			6.12 (2.45)			6.24 (2.47)		
Bayes, 5 Actives	Number of Features		all	200	500	1000	200	500	1000	200	500	1000	200	500	1000	
	0.5/Å ²		2.34 (0.38)	4.02 (1.92)	3.96 (1.52)	4.00 (1.62)	4.96 (2.00)	4.98 (1.50)	5.38 (1.24)	5.86 (2.14)	5.64 (2.44)	4.78 (2.06)	7.16 (1.64)	6.64 (2.00)	6.50 (2.28)	
	2.0/Å ²		1.36 (0.80)	4.08 (1.74)	3.44 (1.44)	2.64 (1.66)	4.84 (1.72)	4.00 (1.56)	4.08 (1.92)	6.08 (1.32)	5.68 (1.44)	4.68 (1.08)	4.60 (1.98)	5.04 (2.72)	4.36 (1.86)	

A comparison of the performance of Tanimoto coefficients and the Naïve Bayesian Classifier is given in Table 22. The point density is varied between $0.5/\text{\AA}^2$ and $2.0/\text{\AA}^2$ and, in the case of the Bayesian Classifier, the number of selected features is varied as well. Given the fact that the Tanimoto coefficient uses single queries and no information from inactive structures, it performs surprisingly well. If at least points adjacent to the central surface point are used for descriptor generation, average hit rates of the Tanimoto coefficient are between 5.68 and 6.08, compared to hit rates between 3.96 and 7.16 for the Bayesian Classifier using 5 active structures (all at $0.5/\text{\AA}^2$). At a higher point density of $2.0/\text{\AA}^2$, average hit rates using the Tanimoto coefficient are between 5.32 and 6.24, compared to between 2.64 and 5.04 if the Naïve Bayesian Classifier is used. The Tanimoto coefficient and the Bayesian classifier thus show opposite tendencies with respect to classification performance if the surface point density is increased (Table 22): Tanimoto performance slightly improves with denser surface points while performance of the Bayesian Classifier decreases. This may be due to assumptions underlying the Naïve Bayesian Classifier employed here, which is the independence of features. If highly correlated features are present in a given molecule that e.g. classifies a molecule to be active, they are all treated as independent features. Classification of the molecule is thus skewed, because the Naïve Bayesian Classifier treats them as independent biases towards activity, although they (in an extreme case) all confer the same information. This result that the Naïve Bayesian Classifier does not perform particularly well in case of partially correlated features (as it is the case here) was also found earlier²³². It is still surprising that, overall, performance of single structures with the Tanimoto coefficient is of comparable performance to the Bayesian Classifier, although the latter method has knowledge about multiple active structures as well as about inactive structures.

Overall performance is compared to other methods in Figure 27. Compared are atom environments⁷⁹ with the Tanimoto coefficient⁸⁰, Feature Trees⁷¹, surface fingerprints with the Bayesian classifier (as described in this work), atom environments with the Naïve Bayesian Classifier⁷⁹, ISIS MOLSKEYS²³⁹, surface point environments with the Tanimoto coefficient (as described here), Daylight fingerprints⁸⁶, SYBYL Hologram QSAR⁸⁷, and three virtual affinity fingerprint methods: Flexsim-X¹³⁰, Flexsim-S²⁴⁰ and DOCKSIM¹²⁹. Performance of methods other than atom environments and surface point environments are taken from Briem and Lessel¹³². The method presented here outperforms the (3D) virtual affinity fingerprint methods

as well as the (2D) Daylight and SYBYL Hologram QSAR fingerprints. One of the reasons for that may be conformational tolerance of this descriptor, as discussed in detail below. Other 3D descriptors, which employ overall distance information between pharmacophores (be it surface points or atom centred pharmacophores) change considerably if the descriptor is calculated for multiple conformations, while this descriptor is reasonably tolerant to conformational changes.

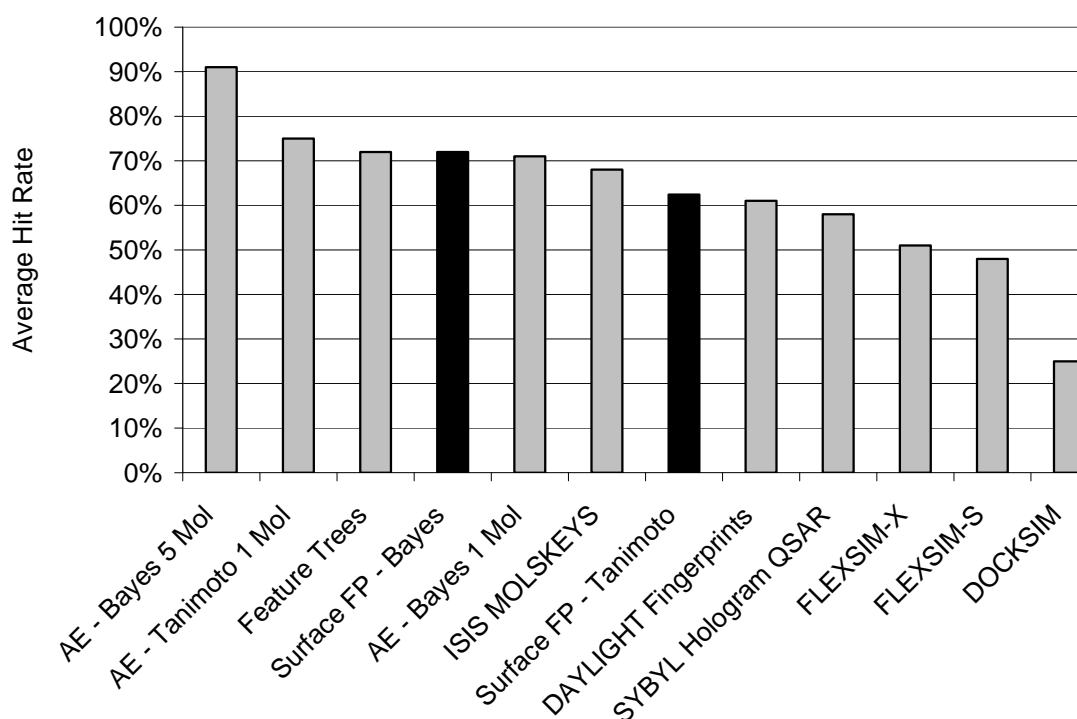


Figure 27. Comparison of similarity searching performance of the surface fingerprint descriptor in combination with other similarity searching methods. Performance of the surface fingerprint descriptor on this dataset is comparable to that of 2D methods. The Bayesian Classifier combines information from multiple molecules and is able to increase classification performance.

Three-dimensional descriptors always depend (to a varying degree) on the particular conformation of the molecule to be described; hence tolerance of the descriptor presented here with respect to conformational changes was examined. Ten molecules from each of the five sets of active compounds were chosen randomly. Using the genetic algorithm conformational search in Sybyl⁸⁷ a set of 10 random conformations of each molecule was created. Genetic search was favoured over the random search option because random searches do not cover conformational space sufficiently well if only a small number of conformations is created. The window size for the genetic search was set to 10° in case of rigid 5HT3 ligands and 100° in case of all other

datasets (ACE, HMG, PAF, TXA2), giving highly diverse conformations. All 10 conformations were put into the database containing “inactive” structures as well as all active structures from the five active datasets, excluding the query structure. All structures of the database were ranked according to Tanimoto similarity to the query structure. For a truly conformationally invariant descriptor, all ten conformations should occur at the top of the sorted list because all descriptors were calculated for different conformations of the same structure. For a very sensitive descriptor, considerable spread throughout the database is expected. The number of different conformations of the query structure among the top 10, 20, 30, 40 and 50 positions of the sorted library was calculated to gauge conformational tolerance of the descriptor.

Table 23. Percentage of conformations of the same structure found at the top of the sorted database. In the top 50 positions (corresponding to about 5% of the database) 94% of the conformations are found.

Percentage of Conformations (out of 10) of the Same Structure Found at Top n Positions of the Whole Database						
Top ... Positions	5HT3	ACE	HMG	PAF	TXA2	Average
10	70	69	75	56	50	64
20	85	87	91	81	70	82.8
30	89	94	94	90	78	89
40	90	96	96	93	88	92.6
50	90	97	96	95	92	94

The influence of conformational variance on descriptor generation is given in Table 23. Nearly two thirds (64%) of all conformations of the same molecule are identified as most similar by the Tanimoto coefficient (placed at the top 10 positions of the sorted list). 94% of all conformations are found in the top 50 positions (roughly 5%) of the sorted library. Thus, if a molecule that is similar to the query molecule is present in the database, it is likely to be ranked at the top of the sorted database. This leads to the tentative conclusion (based on the five different datasets and diverse sets of conformations employed here) that the descriptor is unlikely to miss an active

molecule when it is just not present in the “correct” conformation (e.g. the binding conformation or any other pharmacophoric conformation) in the database.

In addition to not being too sensitive to conformational changes, also different rotations of the molecule should give very similar (in case of truly invariant descriptors identical) features for every orientation of the molecule in space. The influence of rotations on descriptor generation was investigated in the following. 10 structures were randomly selected from the dataset and each of them was saved in two arbitrary rotations. Next, descriptors for both molecular orientations were calculated, employing all probes and four layers adjacent to the central surface point. The molecule was then compared to itself (given its two different spatial orientations), and to the rest of the database molecules. The result of this calculation is shown in Figure 28 below, showing the relative frequency of Tanimoto similarities for both cases (molecule against itself and molecule against library) with their Gauss-representations.

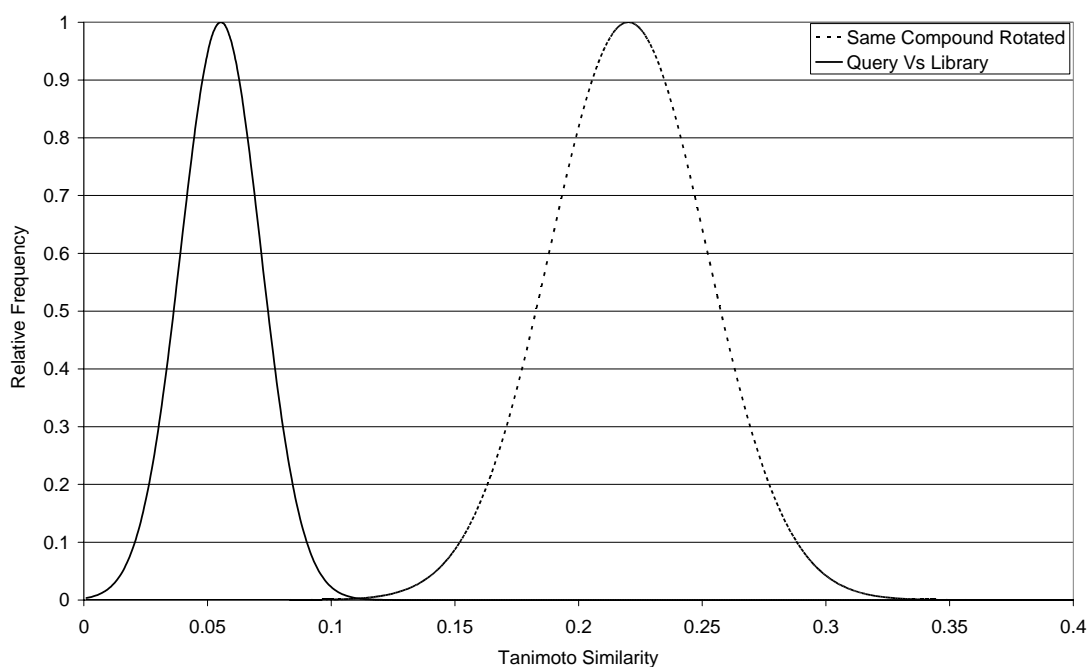


Figure 28. Tanimoto similarity distribution of compounds against themselves in random spatial orientations and against the remaining database of compounds. While the descriptor is not invariant in an absolute sense (the coefficient does not approach one), self-similarity is considerably larger (by five to seven standard deviations) than similarity to other database molecules.

Several conclusions can be drawn from this graph. The first one is that, in absolute terms, the descriptor indeed *does* depend on the particular molecular orientation since

Tanimoto similarity of 1 is not achieved in any of the cases. The second conclusion is that the scores of the descriptor employed here are generally much smaller than in case of two-dimensional descriptors, for example circular fingerprints, even more so structural key fingerprints. Tanimoto scores are around 0.05 in the case here instead of several times as large in case of other descriptors (e.g. around 0.3 in a theoretical study¹⁶⁰). This can be attributed to the large possible number of features: Given that 4 layers in combination with 8 bits each are employed to encode an individual feature, 2^{40} or roughly 10^{12} features can be constructed. Matching features are therefore a rather rare event, leading to, on average, smaller scores. By putting the two distributions shown in Figure 28 in relation to each other, it can be observed that, while in *absolute* terms the descriptor is not rotationally invariant, the scores of two structures modified by rotation are still *significantly larger* than the scores obtained for the whole database. The mean score for rotationally transformed molecules is about seven standard deviations away from the score of the average library molecule in standard deviations of the whole-library distribution. Express as standard deviations of the scores of translationally rotated structures the mean scores are five standard deviations apart. To summary, the descriptor presented here that is not rotationally invariant (since not exactly the same descriptor is produced for different rotations), but it is rotationally *tolerant enough* to give highest scores for identical (but rotated) molecules *on a relative scale*.

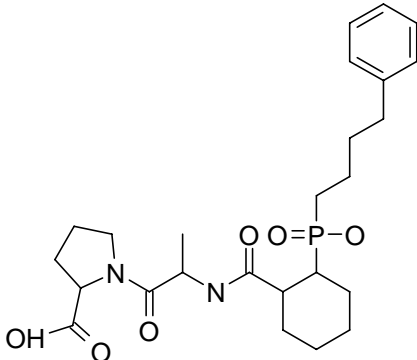
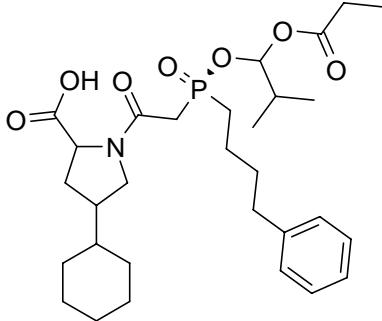
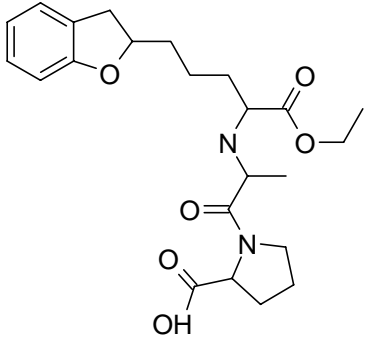
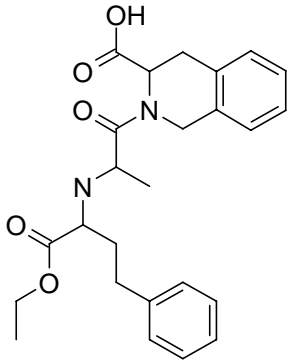
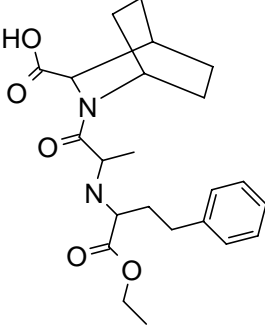
In addition to finding active structures (examined in the preceding calculations) it is one of the superior properties of 3D descriptors over 2D descriptors that they potentially facilitate “scaffold hopping”; the finding of structures which possess shape and pharmacophore similarity without being similar with respect to their connectivity tables. This is illustrated using one query from the data set of ACE inhibitors (Table 24) and the data set of Thromboxane A2 antagonists (Table 25).

Table 21 shows the query (ACE inhibitor) used to screen the database and the highest ranked structures found. Of the 10 most similar compounds retrieved all except nos. 6,7 and 10 are classified as being ACE inhibitors in the MDDR database. (One might think that one only needs to identify amide bonds or carboxylic acids to identify ACE inhibitors, but this is not sufficient since there were more than 100 structures in the database containing either of those features without being classified as an ACE inhibitor). To gauge complementarity of our method to established methods, the same query was used to screen the database using seven other methods implemented in

MOE²⁵²: MACCS Keys, 2D-graph based 3-point pharmacophores (GpiDAPH3), typed atom distances (TAD), typed atom triangles (TAT), typed graph triangles (TGT), 3D 3-point pharmacophores (piDAPH3) and 3D 4-point pharmacophores (piDAPH3). The total number of retrieved structures varied between 2 (typed atom distances) and 8 (MACCS keys). Only surface point environments, graph-based 3-point pharmacophores and MACCS keys retrieved structures that were not retrieved by any other method – 5, 1 and 4 additional compounds, respectively. Five of the active structures found by our method (no. 3, 4, 5, 8 and 9 in Table 25) were not found by any of the other seven methods employed.

It should be acknowledged here that a more systematic analysis of the overlap of retrieved compounds should be performed in the future.

Table 24. Query (ACE inhibitor) used to screen the database and the highest ranked structures found (out of which all except no. 6,7 and 10 are classified as being ACE inhibitors in the MDDR database). Five of the active structures found (no. 3, 4, 5, 8 and 9) were not found by any of the other seven methods employed.

Query			
Ranking position	Structure	Ranking position	Structure
1		2	
3		4	

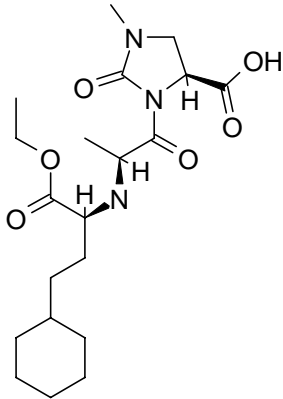
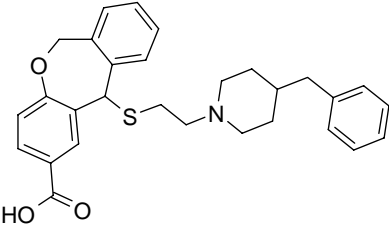
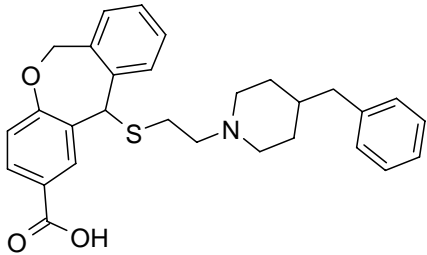
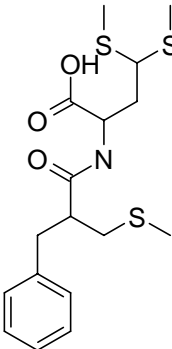
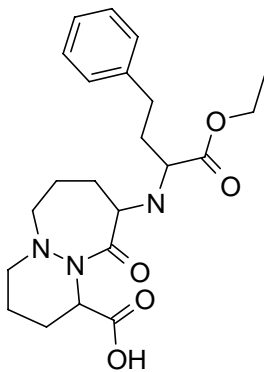
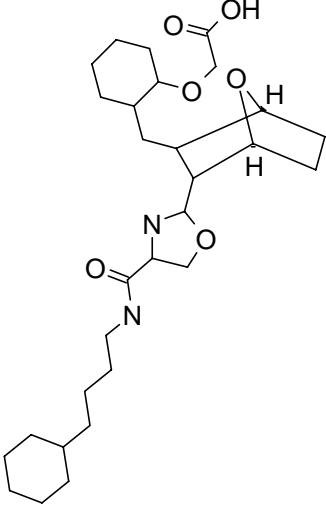
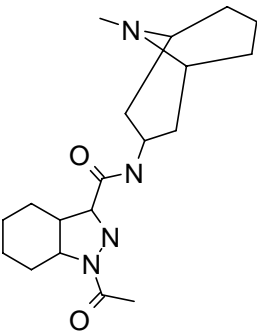
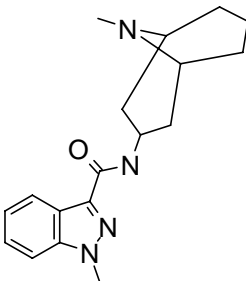
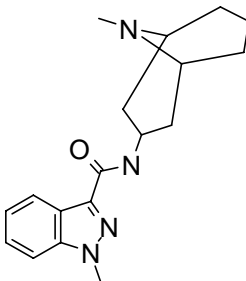
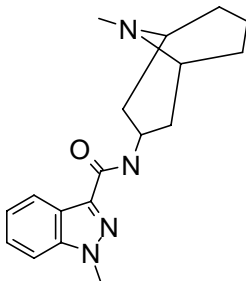
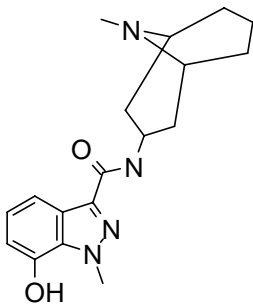
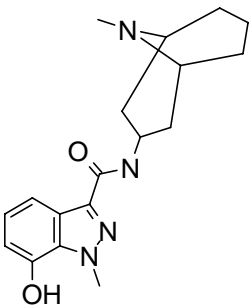
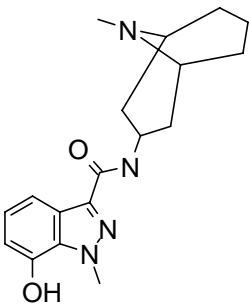
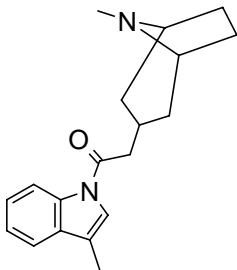
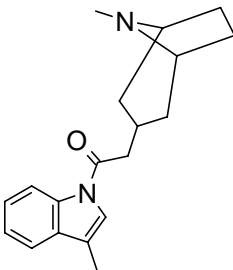
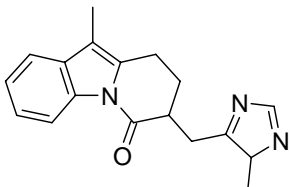
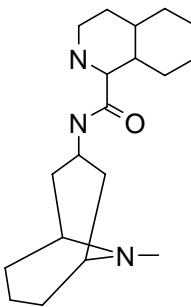
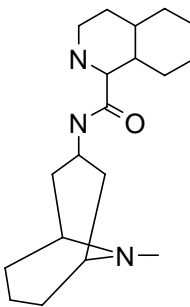
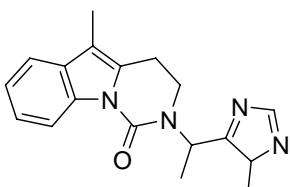
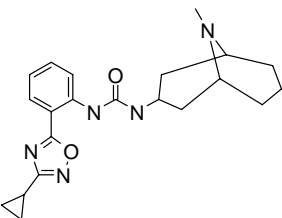
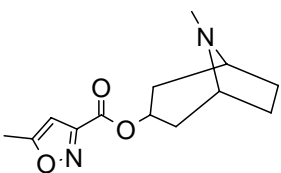
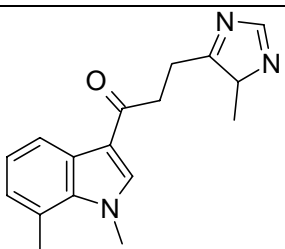
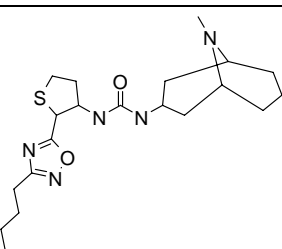
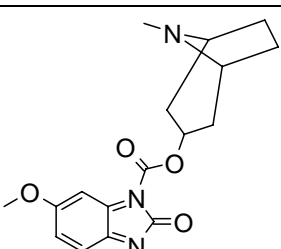
5		6	
7		8	
9		10	

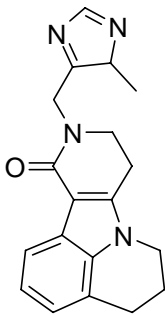
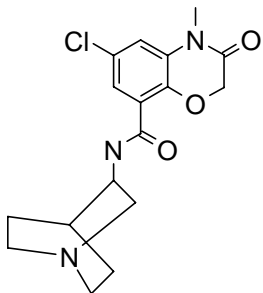
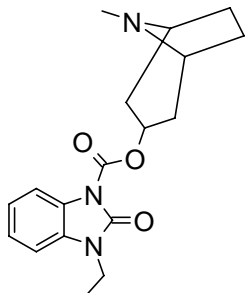
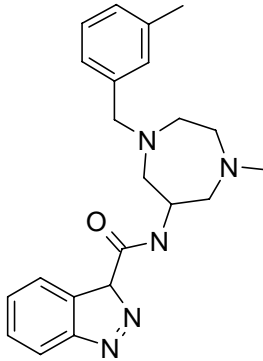
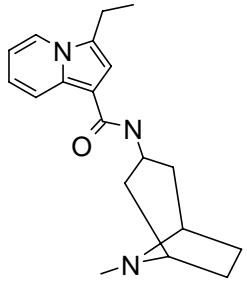
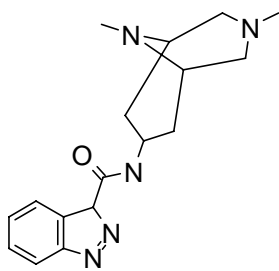
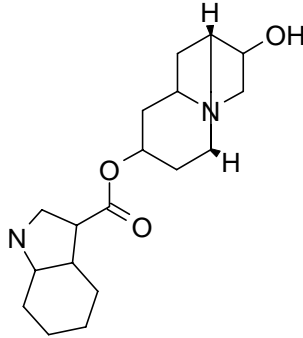
Table 25 compares the active structures found using a Thromboxane A2 antagonist query and different similarity searching methods. Compared are surface point environments, spatial three-point pharmacophores (TAT) and graph-based 3-point pharmacophores (GpiDAPH3, as implemented in MOE). The total number of active structures retrieved is 7 for surface point environments, 7 for spatial 3-point pharmacophore and 10 for graph-based 3-point pharmacophores. While graph-based 3-point pharmacophores retrieve the highest number of active structures, all except

one of the structures contain the bicyclic ring system also present in the query compound. In contrast, surface point environments retrieve only seven active compounds among the 10 most similar structures, but on the other hand 4 out of the 7 active compounds retrieved do not retain the bicyclic ring system of the query compound. In addition, three different scaffolds are present among this subset of retrieved active compounds without the ring system. Illustrated by two examples using an ACE inhibitor and a TXA2 antagonist as query compounds, the method presented here seems to complement established 2D and 3D methods used currently for similarity searching.

Table 25. Query (TXA2 antagonist) used to screen the database and active structures found among the ten most similar compounds. Compared are surface point environments, graph-based three-point pharmacophores and spatial three-point pharmacophores. While other methods retrieve in total more active structures, the method presented here retrieves more dissimilar structures to the query in this example.

		
Surface Point Environments	Typed Atom Triangles (TAT)	Graph-Based Three-point-pharmacophores (GpiDAPH3)
Identical Active Structures Found (all with bicyclic system / 2D-similar to query)		
		

Surface Environments	Point	Typed Atom Triangles (TAT)	Graph-Based Three-point-pharmacophores (GpiDAPH3)
			
			
Other structures found which are less 2D-similar to the query			
			
			
			

Surface Environments	Point	Typed Atom Triangles (TAT)	Graph-Based Three-point-pharmacophores (GpiDAPH3)
			
n/a			
n/a	n/a		
n/a	n/a		

Finally it is likely that a method captures sensible features for classification (as opposed to randomly finding active compounds) if it performs well on several different data sets. By selecting features which are identified as being important for

activity by the algorithm and projecting them back on the molecular surface, it can be verified that they do not constitute incomprehensible sets of features which are only accidentally correlated with activity. (In most data sets variables like this exist, which enable classification but are still meaningless.) In addition, the projection of features on the molecular surface may provide insight into ligand features responsible for binding. This is illustrated by projecting features of inhibitors of 3-hydroxy-3-methylglutaryl coenzyme A reductase, Angiotensin Converting Enzyme and features of antagonists of Thromboxane A2 back onto the molecular surface. Surface fingerprint descriptors were calculated at point densities of $2.0/\text{\AA}^2$ for all six interaction fields and using layers 0 – 4 for descriptor generation. Information gain feature selection was performed to select those features possessing highest information gain, which were more frequent in the set of active molecules. Those features are shown in Figures 29, 30 and 31. Figure 29 shows features selected to be characteristic for a 3-hydroxyl-3-methylglutaryl coenzyme A reductase inhibitor, Figure 30 shows features from an Angiotensin Converting Enzyme inhibitor and Figure 31 illustrates the selected features using a Thromboxane A2 antagonist.

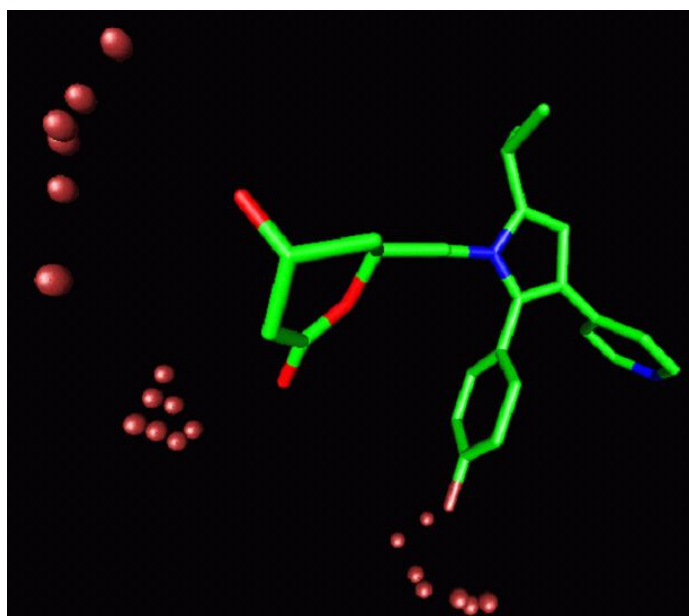


Figure 29. Features identifying the putative pharmacophore of a 3-hydroxyl-3-methylglutaryl-CoenzymeA reductase inhibitor. The polar interactions in the upper left corner and the lipophilic interaction of the fluorobenzyl moiety match binding patterns observed in crystal structures of other HMG-CoA inhibitors. While the compound shown is indeed a prodrug, characteristic features conferring activity are nonetheless identified correctly.

Selected features of the HMG-CoA inhibitor in Figure 29 are adjacent to oxygen substituents on the left hand side and the lipophilic ring at the bottom of the figure. Crystal structures of HMG-CoA reductase complexed with statins¹³⁹ show a common binding pattern between the carboxylic acid and hydroxyl groups of the HMG moiety and polar sidechains of the protein. In addition a lipophilic cleft perpendicular to the axis of polar interactions is present, which is surrounded by a flexible α helix that is able to accommodate lipophilic groups of different shapes and sizes. Both features, the oxygen atoms corresponding to the polar interactions of the HMG moiety and the lipophilic fluorobenzene, are correctly identified by the algorithm.

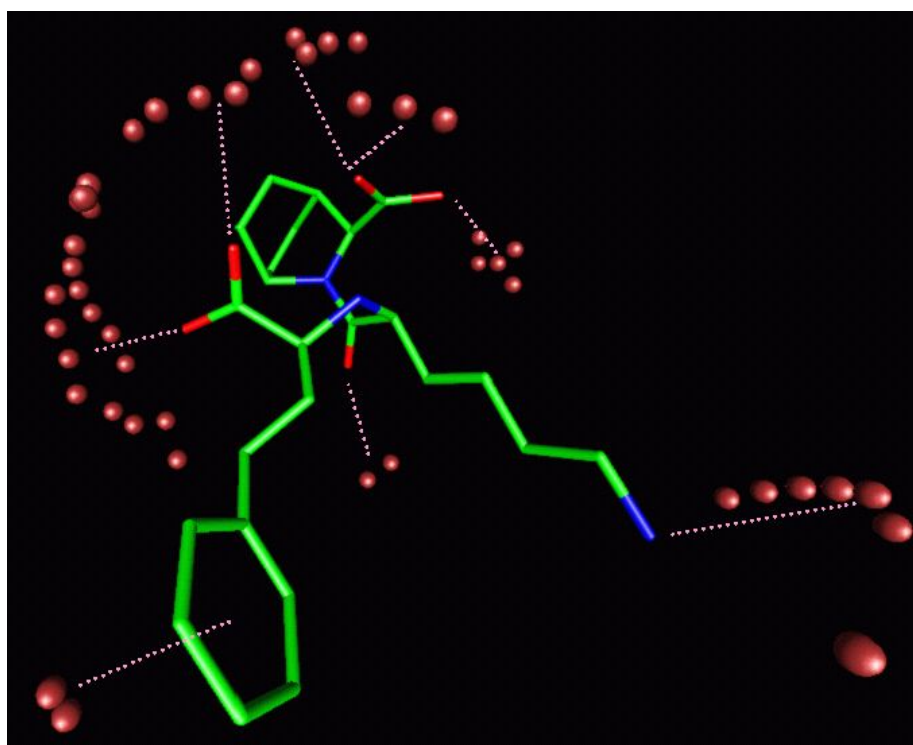


Figure 30. Features identifying the putative pharmacophore of an Angiotensin Converting Enzyme inhibitor. Both lipophilic interactions (the aromatic ring in the lower left corner and hydrogen bonding and charge interactions (in the upper left hand corner and in the top middle of the figure) are identified by the algorithm, based solely on ligand information.

Selected features of the ACE inhibitor in Figure 30 are assigned to various carbonyl oxygens and lipophilic moieties. The experimentally determined binding site of ACE^{236,238} exhibits pairs of hydrogen bond donors and hydrogen bond acceptors as well as lipophilic pockets. The algorithm identifies lipophilic rings and hydrogen bond accepting carbonyl groups as well as the carboxylic acid, which was thought to interact with a bound Zn^{2+} in the enzyme. Although deemed to be important and

successfully used for the design of ACE inhibitors²³⁶, the recently resolved crystal structure of an Angiotensin Converting Enzyme/lisinopril complex did not show a zinc binding site interacting with this ligand²³⁸.

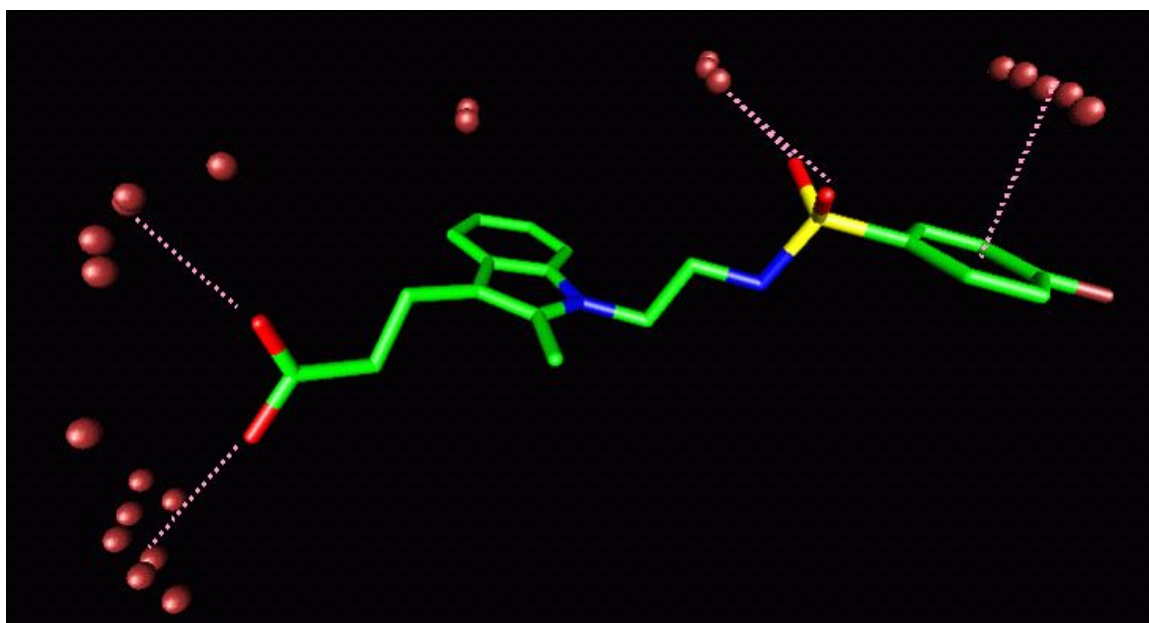


Figure 31. Features identifying the putative pharmacophore of a Thromboxane A2 antagonist. Polar interactions of the carboxylic acid group on the left hand side, hydrogen bond acceptor potential of the sulfonamide moiety and the lipophilic interaction of the fluorobenzyl ring match binding patterns derived from homology models of the binding site. The bound conformation of the ligand is likely to be bent²⁵³ at an angle of about 90 degrees so that the lipophilic rings points downward.

Binding of the TXA2 antagonist in Figure 31 is suggested to be enhanced by interactions of the carboxylic acid on the left hand side, aromatic interactions and hydrogen bond acceptor properties in the centre of the figure and a fluoro-substituted benzene ring, shown on the right hand side of the figure. This binding pattern can be compared to a ligand-target complex derived by homology modeling²⁵³. An arginine residue of Thromboxane A2 is thought to form a charge interaction with a carboxylic acid group of the ligand. A serine residue from the target in this model forms a hydrogen bond interaction, where a hydroxyl group of the ligand acts as an acceptor. In addition, a large lipophilic pocket is present perpendicular to the arginine-serine axis. All three features, carboxylic acid hydrogen bond acceptor (in this case a sulfonamide group) and the fluorobenzene which points in the lipophilic pocket, are identified correctly by the algorithm presented here. This is achieved without having the binding conformation of the ligand available, which is more likely to be bent (the

C-C bond between the sulfonamide and indole moiety can be rotated by 180° at both carbon atoms to achieve a bent conformation).

Back-projection of the features deemed to be responsible for binding on different molecular scaffolds should also be able to identify fragments which are similar with respect to their binding properties, despite showing different atom types and / or connectivity (bioisosteres). An example of that capability is shown in Figure 32 for two compounds having antagonistic properties on TXA₂. Lipophilic interactions are formed *via* an aliphatic chain in case of the first compound and *via* a halide-substituted benzyl ring in case of the second compound. Hydrogen bonds are formed *via* a cyclic ether in the first case and *via* a sulfonamide in the second case. Charge interactions require a carboxylic acid function in both cases.

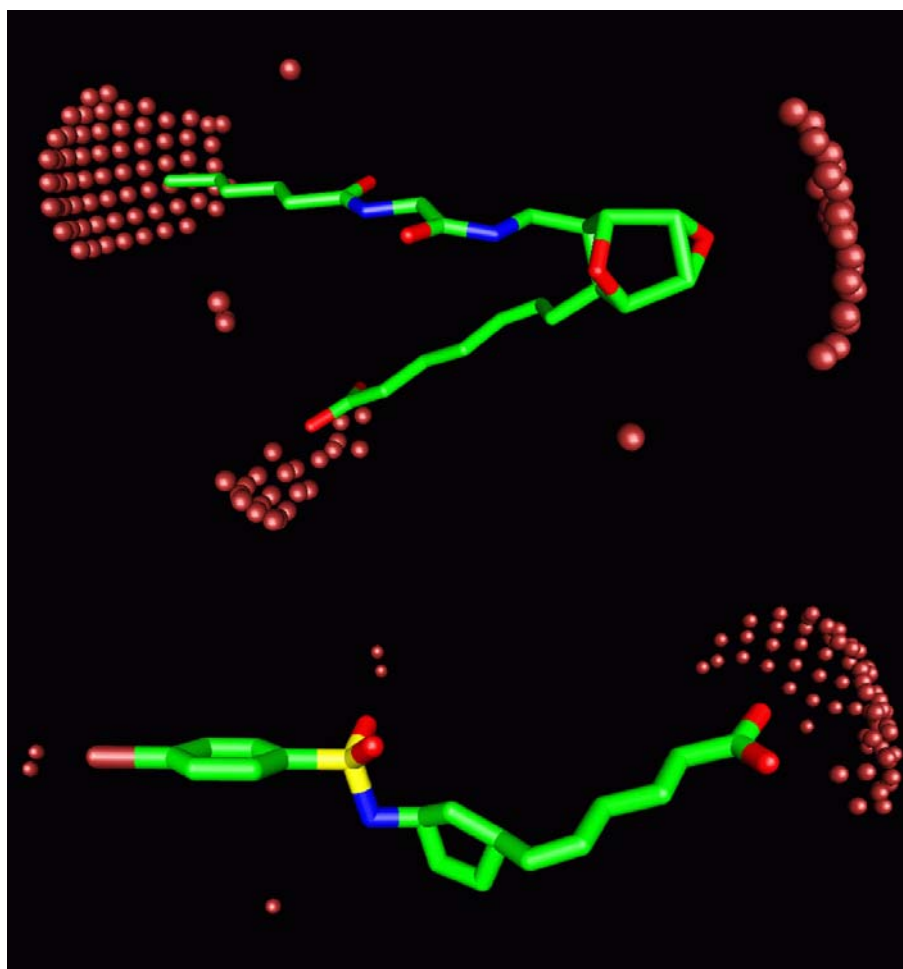


Figure 32. Features identifying the putative pharmacophore of a Thromboxane A₂ antagonist, back-projected onto two active compounds with different scaffolds. Lipophilic interactions are mediated by an aliphatic chain in the first case and by a halide-substituted benzyl ring in the second case. Hydrogen bonds are formed through a cyclic ether in the first case and a sulfonamide in the second case.

Overall, the features selected from the information-gain based feature selection exhibit similar binding patterns to those observed experimentally or in modeling studies. This is achieved without knowledge of the target nor about the conformation of the ligand in the bound state.

Finally, we would like to comment on the use of local information for descriptor generation. Every descriptor derived from 3D coordinates depends on the particular conformation of the molecule described. There exists a considerable trade-off: the more focused local descriptors do not include any distance information, but are on the other hand invariant to conformational changes. Descriptors that include inter-descriptor distance information potentially cover all possible pharmacophore point combinations but are on the other hand very dependent on the particular conformation chosen. It is likely that an ‘optimum’ descriptor for a particular task lies between those extremes.

To illustrate the dependence of intra-molecular distances on conformation, the ACE inhibitor from Figure 33 was subject to a 10ps molecular dynamics simulation *in-vacuo* using Sybyl. Standard settings were used in combination with a NTV ensemble at 310K, the Tripos force field, Gasteiger-Huckel charges and a distance-dependent dielectric function. The distance between the outer carbon of the aromatic ring and the amide oxygen, the nitrogen in the alkyl chain and the oxygen atoms of the closer and terminal carboxylic acid groups were recorded for the run time of the simulation (illustrated in Figure 33). The time-dependent distribution function of intramolecular separations is given in Figure 34. While the intra-feature distance between close features such as the aromatic ring and the amide oxygen shows a sharp peak, the intra-feature separation between all others shows considerable variation. The distance between the aromatic ring and the distal carboxylic acid moiety varies between 7Å and 18Å with a “forbidden zone” between 12Å and 16Å. This illustrates conformational problems when distances between features are taken into account.

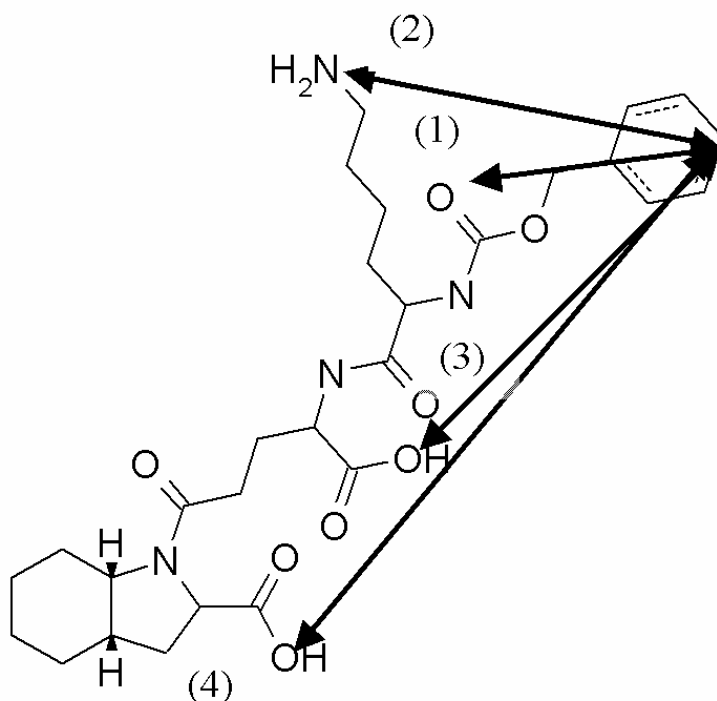


Figure 33. Illustration of the distances measured during the MD simulation of an ACE inhibitor. All distances recorded are taken from the outer atom of the aromatic ring to heteroatoms throughout the rest of the structure.

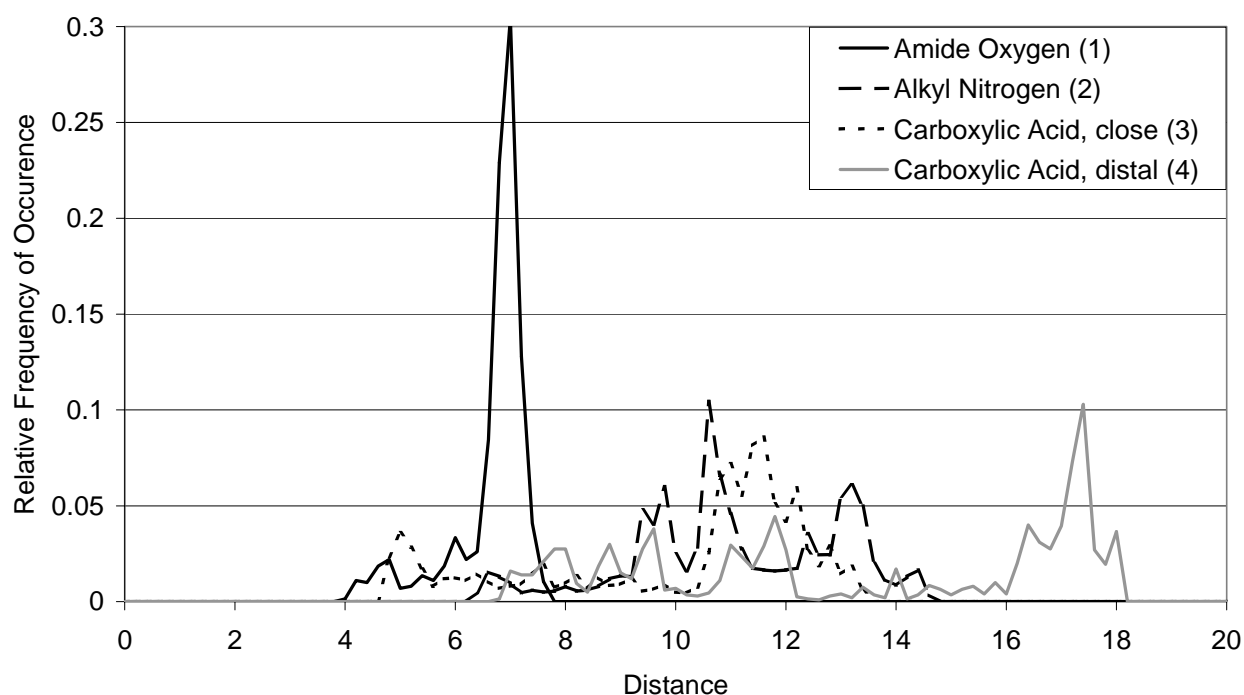


Figure 34. Relative frequencies of distances between the outer carbon of the aromatic ring and the amide oxygen, alkyl nitrogen and two carboxylic acid groups of the ACE inhibitor shown in Figure 33. Distances are derived from a 10ps molecular dynamics simulation in Sybyl and illustrate that distances between features depend to a great extent on the particular conformation chosen.

Conformational analysis of this type will help identify lower energy conformations, particularly where the molecules are predominately rigid, therefore minimizing conformational flexibility. However in most cases, the receptor bound conformation is not the *in-vacuo* (or solvated) lowest energy conformation. In a recent study²⁵⁴, it was found that 60% of the ligands studied do not bind in a local minimum conformation. Strain energies of at least 9 kcal/mol were found in more than 10% of the bound ligand conformations. Including conformational constraints is often difficult in the absence of experimental or other evidence (e.g. using the active analogue approach¹⁰⁷) of the receptor bound conformation.

The surface point environment descriptor introduced here (coupled with feature selection and a Naïve Bayesian Classifier) represents part of the molecular surface, spanning about 8Å in diameter and seems reasonably tolerant to conformational changes (in small drug-like molecules) when used for database searching. Also it is important to note that the classification of a molecule generally depends on more than one feature (in practice usually several hundreds). Some features may overlap and therefore they will represent continuous regions considerably greater than 8Å in diameter. Probability of class membership will thus depend on a number of different features, in effect representing an implicit AND of those features. Chosen features may represent continuous or discontinuous regions, allowing flexible representation of important field properties.

8.3 Conclusions

We present a novel similarity searching algorithm based on surface point environment descriptors in combination with the Tanimoto coefficient and the Naïve Bayesian Classifier. It shows high retrieval rates, the identification of active structures with different scaffolds and back-projectability of features which can be correlated with experimentally determined binding patterns. Used in combination with Tanimoto coefficients, its performance is comparable to that of commonly used 2D fingerprints. If the Tanimoto coefficient is replaced by a Bayesian Classifier, information from multiple structures can be combined.

The descriptor is shown to be tolerant to conformational variations of the ligand structure. On average, two thirds of randomly generated conformations of sets of ten structures each from five activity classes are classified as being most similar to one conformation used for querying the whole database. This database in this case

contains more than 900 structures in total and 39-131 structures of the same activity class (depending on the particular class chosen). This implies that in most cases, active structures were not missed where conformations of similar molecules present in the database were not the ones which would give the best (conformational) match to the query.

Active structures retrieved by this approach possess a variety of scaffolds, as illustrated by a database search using an ACE inhibitor and a Thromboxane A2 antagonist. Active structures with no apparent 2D similarity to the query are identified, showing that the method is capable of “scaffold hopping”. The active compounds retrieved (as illustrated by those two cases) were also not found using seven other 2D and 3D similarity searching methods. This indicates complementarity of the algorithm presented here to established similarity searching algorithms, although a systematic comparison of compounds retrieved would be clearly of additional value.

Feature selection is shown to identify important features which, if projected back on to the molecular surface, can be associated with experimentally observed binding patterns. This is achieved without alignment of structures or information about the target structure. Illustrations of feature selection are given for inhibitors of 3-hydroxy-3-methylglutaryl coenzyme-A and Angiotensin Converting Enzyme as well as antagonists of Thromboxane A2. Features contributing to the binding of a Thromboxane A2 antagonist are correctly identified, even where the structure is not given in the observed binding conformation. Features responsible for binding are also able to identify bioisosteric fragments of the molecule.

According to the idea that ligand-target interactions are mediated *via* interactions of the two molecular surfaces, localized surface point environments²⁵⁵ have been used in combination with GRID^{35,103} derived energetic molecular surface properties for similarity searching. Several different molecular probes were employed in order to capture areas of the molecular surface which correspond to putative ligand-target interaction types, such as hydrogen bond donors and acceptors, positively and negatively charged surface areas and lipophilic moieties.

Still, force-field (such as GRID) derived descriptors possess several serious shortcomings. A number of different probes need to be used to ensure that different interaction types are covered sufficiently. This increases the time needed for descriptor generation as well as introducing a degree of arbitrariness into the choice of

probes. Also, most force fields (with exceptions such as the Cresset XED force field²⁵⁶) only employ approximations of molecular properties, such as point charges which do not account for the directionality of lone pairs. They also do not capture sufficiently other properties of the electronic structure such as polarizability and they depend on a parameterization performed using a particular (arbitrary) data set.

The sum of those shortcomings led us to believe that a quantum-mechanical method for the description of molecular surfaces may be more appropriate since it eliminates all of the points above, of course bought at higher computational expense. In this work screening charges of the molecular surface are calculated by the COSMO²⁵⁷ methodology and capture potential intermolecular interactions *via* a calculation of screening charges on the molecular surface. COSMO provides a set of surface patches, typically in the order of several thousand patches for a small molecule, with associated surface screening charges. In the COSMO extension to Realistic Solvation (COSMO-RS)²⁵⁸⁻²⁶⁰ the special importance of the surface screening charge density for electrostatic interactions, hydrogen bonding and interactions of lipophilic regions has been elucidated.

The following chapter describes briefly the background of COSMO and COSMO-RS, the encoding scheme used for descriptor generation and the molecular database employed. Results are presented and discussed in the subsequent chapter. Finally, we give our concluding remarks about the performance of the method and envisaged future work.

9 COSMO-derived screening charges and their relevance for molecular interactions

While we have shown in the previous chapter that overlapping ‘local’ surface patches are well able to discriminate between compounds of different activity classes, the surfaces generated were not ideal. For example, multiple probes had to be employed to cover putative interaction types of the ligand with the target exhaustively. Also, force fields do not capture directionality of lone pairs which are able to function as acceptors of hydrogen bonds – which are known to be beneficial to binding only in very narrow distance and angular ranges²⁶¹.

The sum of those shortcomings led us to the belief that a more rigorous quantum-mechanical method would be the method of choice, more precisely the COSMO implicit solvent quantum mechanical model. It calculates effective surface charges in solution which allows the discrimination between different ligand-target interactions (charge, hydrogen bond and lipophilic interactions) on a single scale. The COSMO model captures the directionality of lone pairs as well as addressing secondary effects and discriminating between accessible and inaccessible atoms of a molecule by the calculation of molecular surface areas which are assigned to individual atoms.

The CONductor-like Screening MOdel²⁶⁰ is a very efficient and robust approximation of the dielectric continuum solvation models²⁶² which is available in many quantum chemical programs. Based on a cavity grid of m surface segments it calculates the polarization by the screening charges of dielectric continuum representing the solvent by the electrostatic field and solute from a scaled conductor boundary condition:

$$f(\varepsilon)\underline{\Phi}^{solute} + \underline{A}^{-1}\underline{q} = 0$$

where $\underline{\Phi}^{solute}$ is the vector of the electrostatic solute potential arising on the m segments of the cavity, \underline{q} is the vector of the screening charges on the m segments, \underline{A} is the Coulomb matrix of the surface segments, and $f(\varepsilon)$ is the a scaling factor depending on the dielectric constant of the medium:

$$f(\varepsilon) = \frac{\varepsilon - 1}{\varepsilon + 0.5}$$

The solute potential $\underline{\Phi}^{solute}$ is calculated by quantum chemical methods, where density functional methods have proved to be the most efficient and reliable. Since the screening causes back-polarization of the solute by the solvent, the dielectric

screening has to be taken into account self-consistently in the quantum chemical calculation. If efficiently implemented, as in the DFT calculations in TURBOMOLE^{263,264}, COSMO calculations can be performed with only small computational overhead compared with gas phase calculations, including consistent geometry optimization in the presence of the dielectric solvent.

Starting from a fundamental criticism of the oversimplified dielectric continuum solvation concept, Klamt developed a statistical thermodynamics extension of the COSMO model named COSMO-RS^{260,265}. In this model the interactions of molecules in a liquid phase are expressed as local contact energies of the molecular surfaces. Here, the COSMO screening charge densities σ , which are the surface charge densities resulting from the set of surface charges q , play the key role for the quantification of electrostatic, hydrogen bonding, and hydrophobic/lipophilic interactions. While originally developed and widely validated for environmental and chemical engineering mixture thermodynamics, the value of the σ -based COSMO-RS concept for the quantification of many ADME properties such as solubility, blood-brain barrier penetration, intestinal absorption, and even for pK_a prediction has also been demonstrated^{265,266}. Furthermore, first applications of the COSMO-RS concept to the evaluation of drug receptor binding are currently being developed. Thus COSMO screening charge densities are likely to provide a sound foundation for investigating ligand-target interactions from first principles.

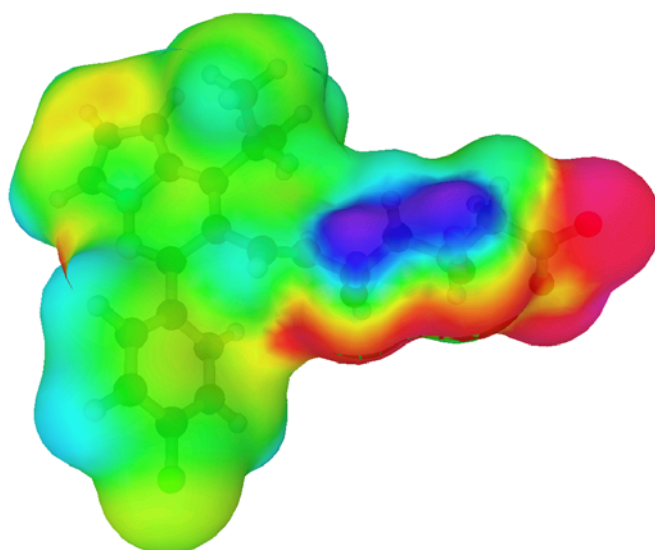


Fig. 35. COSMO-derived screening charge densities σ of an HMG-CoA reductase inhibitor. Hydrogen bond acceptor features as well as negative charges are encoded in red while blue color denotes hydrogen bond donor potential. Lipophilic moieties are colored in green.

The relevance of screening charges calculated by COSMO to molecular binding is illustrated in Figure 35 for an HMG-CoA reductase inhibitor (statin). As known from crystal structures, binding of statins to HMG-CoA reductase is mediated by charge interactions of a carboxylic acid group of the ligand as well as hydrogen bond acceptor functions to the pyruvate-binding site of HMG-CoA. In addition a lipophilic function of the ligand is required which binds to a floppy lipophilic pocket of the target protein. All these features can be well distinguished from the COSMO screening charge densities σ , as illustrated in Figure 35. The carboxylate function is shown on the right in red and purple, while hydrogen bond acceptor functions can readily be identified at the bottom of the same chain. Hydrogen bond donor functions point towards the viewer and are shown in blue while the lipophilic bulk of the structure is given in green colour.

9.1 Material and methods

COSMO screening charge densities σ were encoded as atom-based three-point pharmacophores (3PP)^{110,114}. By projection back on the associated atom centres average σ -values were calculated for each heavy atom and hydrogens attached to elements other than carbon. Atoms with average screening charge densities $\sigma > 0.014 \text{ e}/\text{\AA}^2$ were classified as bearing strongly negative partial charge (type N) and those with average charge densities $0.014 \text{ e}/\text{\AA}^2 \geq \sigma > 0.009 \text{ e}/\text{\AA}^2$ as hydrogen bond donors (D). Negative screening charge densities were associated with atoms showing strongly positive partial charge (P) at $\sigma < -0.014 \text{ e}/\text{\AA}^2$ and hydrogen bond acceptors (A) at $0.014 \text{ e}/\text{\AA}^2 \leq \sigma < 0.009 \text{ e}/\text{\AA}^2$. Atoms with intermediate screening charge densities were classified as lipophilic atoms (L). This results in features broadly in agreement with chemical intuition such as that the doubly bound oxygen of an ester group but not its neighbouring sp^3 hybridized oxygen possesses hydrogen bond acceptor properties. Eight bins were used to encode geometry of the putative pharmacophore triangles, starting at 2 Å and employing bin borders at 3.5, 5, 6.5, 8, 9.5, 11, 13 and 15 Å. Triangles were rotated to a unique orientation before encoding was performed. Triangle counts were kept and molecules were compared using a Tanimoto-like similarity coefficient⁴ which divides the number of matching features by the total number of features present to give a similarity value in the range [0; 1] and also taking into account the size of the structure.

A variety of other encoding schemes were initially employed as well, but since no superior performance (with respect to hit rates) was obtained these routes were not followed further. The range of σ values from $-0.04 \text{ e}/\text{\AA}^2$ up to $+0.04 \text{ e}/\text{\AA}^2$ was binned into 8 (80) equidistant ranges and a frequency histogram for each molecule was obtained, combined with the Euclidean distance as a similarity index. For 8 bins an average hit rate of 4.3 compounds and for 80 bins an average number of 6.0 compounds among the ten nearest neighbours of each active structure was retrieved. Smoothing (i.e. assigning each surface charge half to its actual bin and a quarter each to neighbouring bins) did give very similar results (average hit rates of 5.8 in both cases). Following the encoding by local surface point environments explored earlier²⁵⁵ hit rates of up to 6.2 were obtained if an environment of 5\AA was taken into account for each surface point and the environment was binned in distances of 1\AA each. While these performance levels are comparable to the three-point pharmacophore route employed here, they also – despite using information about the surface charges directly – show no superior performance which was the reason for us to resort to the simpler, atom-centred representation.

For evaluation of the algorithm, 957 ligands extracted from the MDDR database²⁴⁷ were used which were also employed throughout this work. The set¹³² 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 belong to any of these activity classes.

Similarity searching performance was established as the number of structures from the same activity class as the query, as found among its ten nearest neighbours¹³². Similarity searching was repeated ten times and the hit rates obtained were averaged.

The distribution of σ , normalized with respect to the size of the molecules, is shown in Figure 36 averaged over all molecules from each of the five activity classes and a random selection of structures from the MDDR (the “negative” set). Already from this averaged distribution of screening charge densities characteristic differences between the sets are visible which can be assigned to underlying common structural features. For example, statins (dataset HMG) have a lower fraction of very negative screening charge densities, as can be seen in Figure 36. This corresponds to the fact

that positively charged sites (usually protonated nitrogens) are virtually non-existent in this dataset. ACE inhibitors on the other hand frequently possess both positively charged sites (due to the presence of lysine and other basic moieties) as well as negatively charged groups (carboxylate groups), both of which are present in Figure 36 as peaks of very negative and very positive screening charge densities, respectively. While these overall distributions of screening charge densities can be interpreted based on structural molecular features and show commonalities and differences between different classes of active compounds, they still lack a geometrical encoding scheme which merges possible interaction types with their arrangement in space. The geometric encoding scheme employed will be presented next.

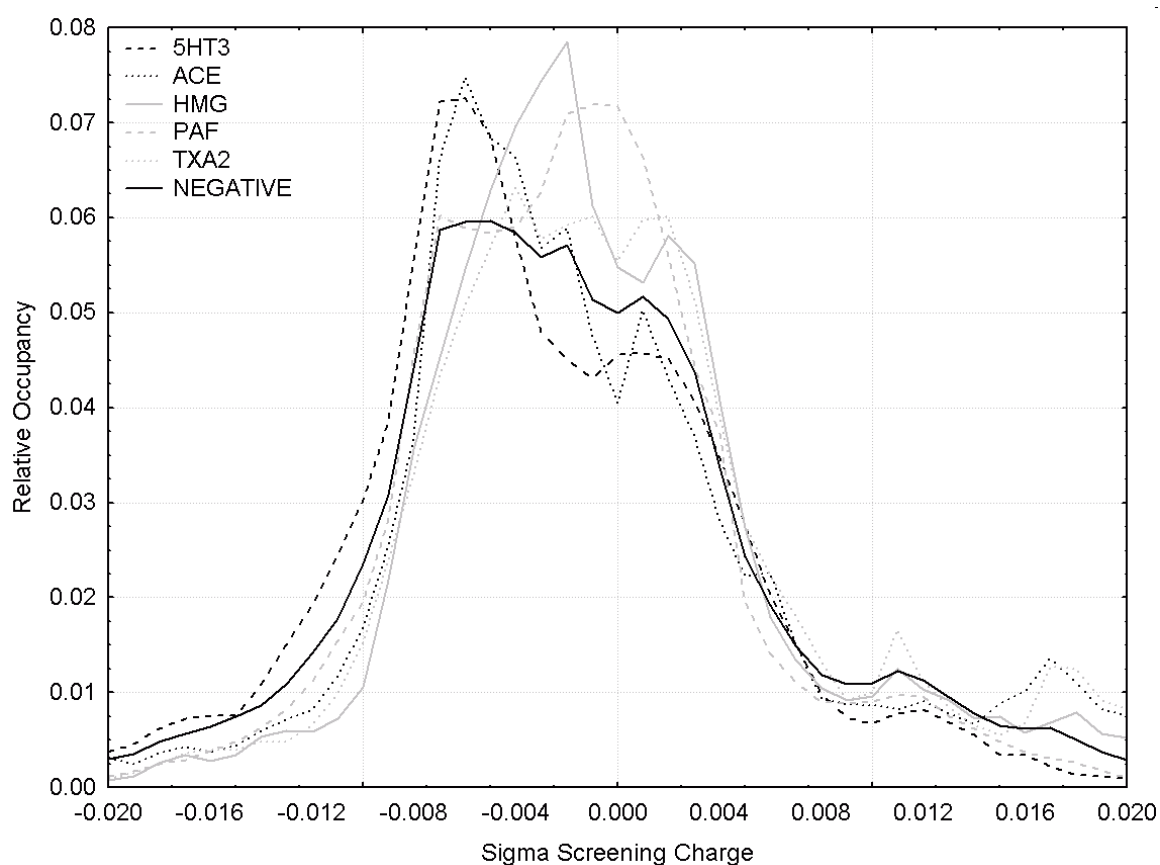


Figure 36. Distributed of surface screening charge densities for the five subsets of active compounds, compared to a random sample of structures from the MDDR (NEGATIVE). For each of the classes characteristic deviations in the sigma-profile from the database average can be observed, rendering this description suitable for the identification of ligands with certain properties from a database.

Structures were exported from the MDDR database in SD format. Protonation states were assigned using MOE²⁵², subsequently 3D structures were generated using CORINA²⁶⁷. Geometries were optimized with AM1/COSMO followed by a single-point BP-SVP-DFT/COSMO calculation using TURBOMOLE²⁶³. Screening charges of surface elements were translated into three-point pharmacophores by a Perl script. Geometry optimization required approximately 10 minutes per compound on a 3 GHz CPU. While we chose COSMO calculations on such a high level for this initial study, it should be noted that a recently developed, very fast screening method COSMOfrag²⁶⁶ can reduce the computational cost to as little as one second per compound if required for large scale screening projects.

9.2 Results and discussion

In the first series of calculations hydrogen atoms bound to hetero atoms were treated as putative pharmacophore points, along with all other hetero atoms of the molecules associated with a solvent-accessible surface area. Hit rates and enrichments obtained from the five classes of active compounds are given in Table 26, together with the associated standard deviations. Considering all hydrogen atoms bound to heavy atoms except carbon as putative pharmacophore points the hit rate for all classes is around 5 (between 4.9 in case of the 5HT3 and TXA2 datasets and 5.2 in case of the HMG dataset), giving an average hit rate of 5.0 and enrichments of about 5- to 14-fold. In the second series of calculations hydrogen atoms bound to hetero atoms were omitted from the list of putative pharmacophore points. In this case performance improves to a mean hit rate of 6.4, corresponding again to enrichments of between 5- and 14-fold. Thus, the introduction of spatial information about the location of putative hydrogen bond donors does not improve performance. Several reasons might be responsible for this finding, among them the fact that ligands often do not bind in their local nor global lowest energy conformation, rendering information about the spatial orientation of functional groups in the liquid phase invalid.

Table 26. Average hit rates, enrichment factors and standard deviations for the five classes of active compounds.

Dataset	All Hetero-H			Hetero-H Except NH		
	Hitrate	Enrichment	σ (Hitrate)	Hitrate	Enrichment	σ (Hitrate)
5HT3	4.9	14.0	3.5	5.6	11.2	3.3
ACE	5.1	7.6	4.1	5.6	13.7	2.8
HMG	5.2	5.8	3.8	9.1	7.9	0.9
PAF	5.1	5.2	3.1	7.0	5.0	1.9
TXA2	4.9	9.4	2.7	4.7	9.4	2.7
Mean	5.0	7.3	3.4	6.4	8.1	2.3

Performance is compared to established methods in Figure 37.

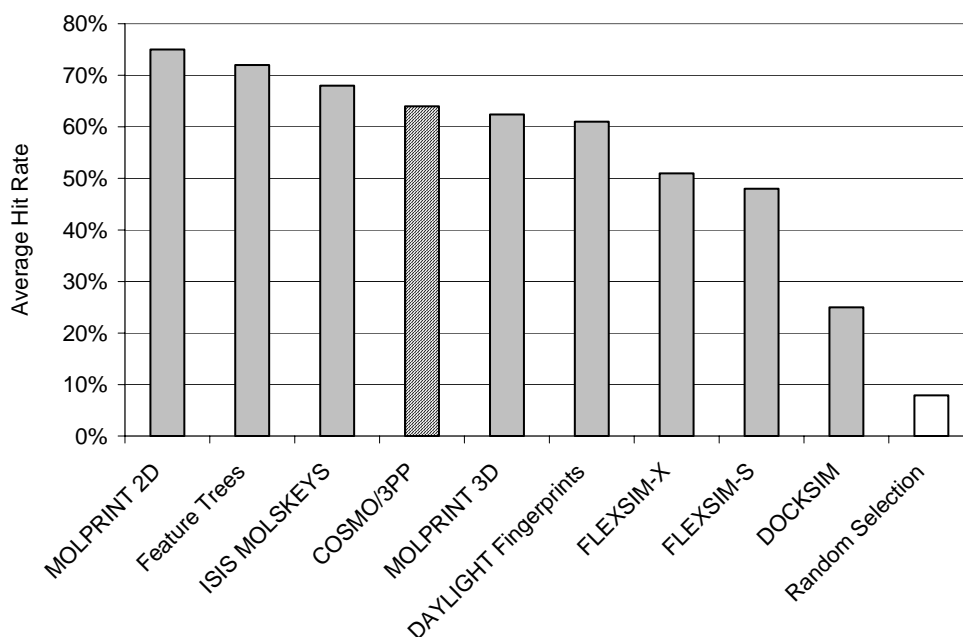


Fig. 37. Comparison of retrieval rates obtained by COSMO-based three-point pharmacophores to established methods. While hit rates achieved are superior to other methods, the added value of the method presented here lies in the type of active structures retrieved.

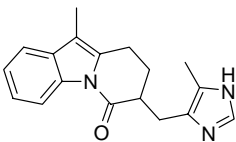
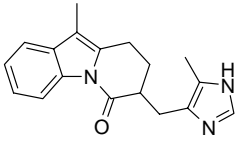
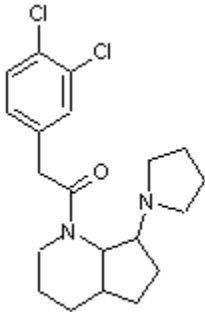
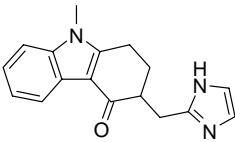
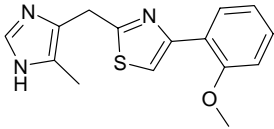
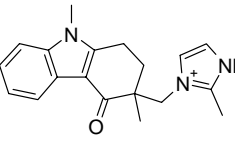
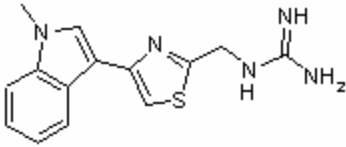
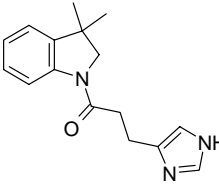
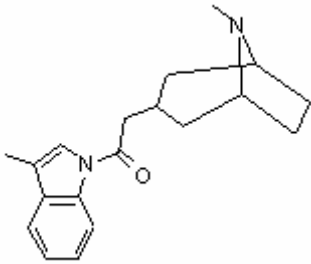
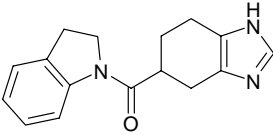
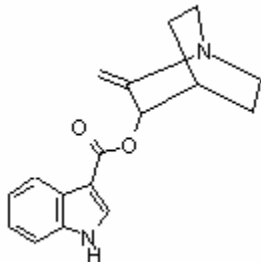
The average hit rate of 6.4 active compounds in the first 10 structures of the ranked list corresponds to slightly better retrieval rates than Daylight fingerprints while not achieving performance of 2D-information based Feature Trees and MOLPRINT 2D

fingerprints. This is in agreement with earlier results on the information content of 2D and 3D descriptors⁸⁴ although it should be noted that the difference remains small and probably not relevant in practice. Nonetheless, we have to state that in the current work we have not yet been able to encode the information contained in the screening charge densities σ in such a way as to improve performance over established methods. This might have on the one hand to do with the restriction of the “molecular similarity principle” mentioned in the introduction. On the other hand, work is still continuing to elucidate an information-preserving way to encode molecular surfaces with associated screening charges, which, as we hope, might lead to improved results.

Still, one advantage of the method presented here over many 2D based methods remains, which is its ability to retrieve active compounds with widely different scaffolds. An example is given in Table 27 for the 5HT3 dataset. If the query shown at the top of the table is used to rank the database of the remaining 956 structures, in this case nine out of the ten most similar compounds retrieved indeed bind to at least one of the Serotonin Receptor subtypes (the datasets contains both agonists and antagonists of all Serotonin Receptor subtypes, thus actually representing a wide variety of activities). Shown are ranking positions with similarity scores and the information whether a compound shows against a 5HT3 receptor subtype.

The compound retrieved at position 1 shows maximum similarity of 1 due to the fact that it is identical to the query and (for an unknown reason) present twice in the MDDR database. While this does not prove any capabilities with respect to similarity searching, this finding shows that the method, despite using three-dimensional structures, is able to generate reproducible descriptors for a given structure. Descending in the list of structures retrieved, similarity to the query continuously decreases. The compounds found at positions 1, 2 and 3 retain the scaffold of the query, while those at positions 4 and 5 retain its amide and heterocyclic moieties. The structures retrieved at positions 7 and 8 already identify novel active scaffolds while those found at positions 9 and 10 of the list display completely novel spiro- and bicyclic ring systems.

Table 27. Structures identified via COSMO screening charge densities σ in combination with a three-point pharmacophore encoding scheme. The query 5HT3 ligand is shown at the top.

			
Rank (T _c)	Structure	Rank (T _c)	Structure
1 (1.00) <i>active</i>		6 (0.56) <i>inact.</i>	
2 (0.66) <i>active</i>		7 (0.54) <i>active</i>	
3 (0.59) <i>active</i>		8 (0.54) <i>active</i>	
4 (0.57) <i>active</i>		9 (0.54) <i>active</i>	
5 (0.56) <i>active</i>		10 (0.51) <i>active</i>	

9.3 Conclusions

We have shown in this work that COSMO screening charge densities σ can successfully be employed for virtual screening and provide a conceptually sound as well as intuitively accessible scheme for calculating putative ligand-receptor interaction types. These interaction types can be represented on a single scale of screening charge densities. Additionally, secondary effects are accounted for as well and consideration is paid to the question of whether atoms are spatially able to interact with the target by possessing solvent-exposed surface area. Retrieval rates are comparable to 2D fingerprints while considerably more diverse compounds are identified. While in this way performance of established methods can be matched, it is also not improved upon, despite the wealth of information available about the molecular surface. Thus, further work will focus on novel encoding schemes which make better use of the screening charge information given; information which is not fully exploited by the atom-based pharmacophores. One example is hydrogen bond acceptor atoms, where the directionality of the lone pairs is well captured by σ , yet neglected in the atom-based encoding scheme. In addition to improving the way information is encoded, the possibility of quantitative activity predictions based on screening charge densities σ will be investigated.

While we chose COSMO calculations on a high level of theory for this initial study, it should be noted that a recently developed, very fast screening method COSMOfrag can reduce the computational cost to as little as one second per compound if required for large scale screening projects. This high-throughput method uses a library of fragments to assemble the approximate screening charge densities σ for a novel molecule. Some technical problems still have to be overcome in order to use COSMOfrag in the context of the COSMO/3PP method described herein.

10 Restrictions of the MOLPRINT 2D circular fingerprint with respect to scaffold hopping, illustrated on the ‘HTS data mining and docking competition’ dataset

While previous chapters were concerned with the development of methods for molecular similarity searching, this chapter describes the application of the presented MOLPRINT 2D method to an independent performance assessment, namely the ‘HTS Data Mining and Docking Competition’. While overall none of the methods presented for HTS data analysis showed satisfactory performance (which were published in the October 2005 issue of the *Journal of Biomolecular Screening*), this also needs to be seen in the context that the test dataset provided was not ideal since it contained only a very small number (if any) valid HTS hits²⁶⁸. Still some lessons for the analysis of HTS can be learned.

10.1 Material and methods

Dihydrofolate reductase (DHFR) catalyzes the reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF). Tetrahydrofolate is essential for the biosynthesis of purines, pyrimidines and some amino acids. This in combination with considerable structural differences between bacterial and human DHFR renders bacterial DHFR an ideal drug target, exploited by highly selective inhibitors such as trimethoprim. Still, development of resistance to trimethoprim creates a need for novel structural series of DHFR inhibitors. As a result a high-throughput screening of 49,995 compounds was recently performed by Zolli-Juran *et al.*²⁶⁹, identifying 32 hits (defined by less than 75% residual activity in both of two screening runs) comprising several novel scaffolds.

The extraction of structural ‘knowledge’ from the compounds and their activities from the first screening (‘training set’) was the goal of the HTS Data Mining and Docking Competition at McMaster University (Ontario). This knowledge was to be used to make predictions about the inhibitory activities of a second set of 50,000 compounds that was to be screened subsequently (‘test set’). Due to the size of the training set the first screening provided a wealth of experimental data for model generation. In order

to exploit this information fully, a sufficiently economical computational method had to be employed.

The descriptor employed by our method, MOLPRINT 2D^{77,79,80}, is based on the connectivity of each heavy atom and its neighbours up to two bonds apart (as described in previous chapters). It is similar to ‘augmented atoms’²²⁸ or the Extended Connectivity Circular Fingerprints (ECFPs) employed by Scitegic²⁰⁸. Feature selection is performed using information-gain based feature selection²³⁰ which was originally devised to induce rules as nodes of decision trees. Ranking employs the Naïve Bayesian Classifier which provides a simple yet surprisingly efficient classification method. Calculation of features, feature selection and ranking are performed at the order of 1000 molecules / second on a 1GHz-PIII Redhat Linux machine, equivalent to response time in the order of tens of seconds on today’s computers on the library size considered here (50,000 structures).

The question of library composition as well as inherent properties of fragment-based descriptors will be revisited in the results and discussion section, which follows the next section, presenting details of the method used.

10.2 Results and discussion

In the first series of calculations a model was built using 32 active structures showing less than 75% residual activity in both runs, which were enriched by 15 other known inhibitors of bacterial and *E. coli* DHFR from public sources. Inclusion of additional compounds was intended to introduce more diversity in the active training set, giving an active set of 47 structures. Since our method is also able to accommodate an inactive dataset, all structures with a residual activity of greater than 100% were selected in the first run to train the inactive model. A residual activity greater than 110% was used to define inactive compounds in the second run, resulting in 32,521 and 4,515 structures respectively. Feature selection was not employed since it did not improve results, possibly due to the broader definition of active and inactive classes if all features are employed for classification. Results of this run are shown in Figure 38. The first 30 active compounds (out of 47; 64% of all actives) are found in the top 200 out of 49,995 positions of the ranked library (0.4%) if all compounds showing residual activity greater 100% are used in the “inactive” set. This corresponds to a – rather hypothetical – enrichment factor of around 160 on the training set. For leave-one out cross-validation results are very similar to the ones shown in Figure 38.

Encouraged by those results this parameterization was used to rank the test set and the results were submitted to the HTS Data Mining and Docking Competition.

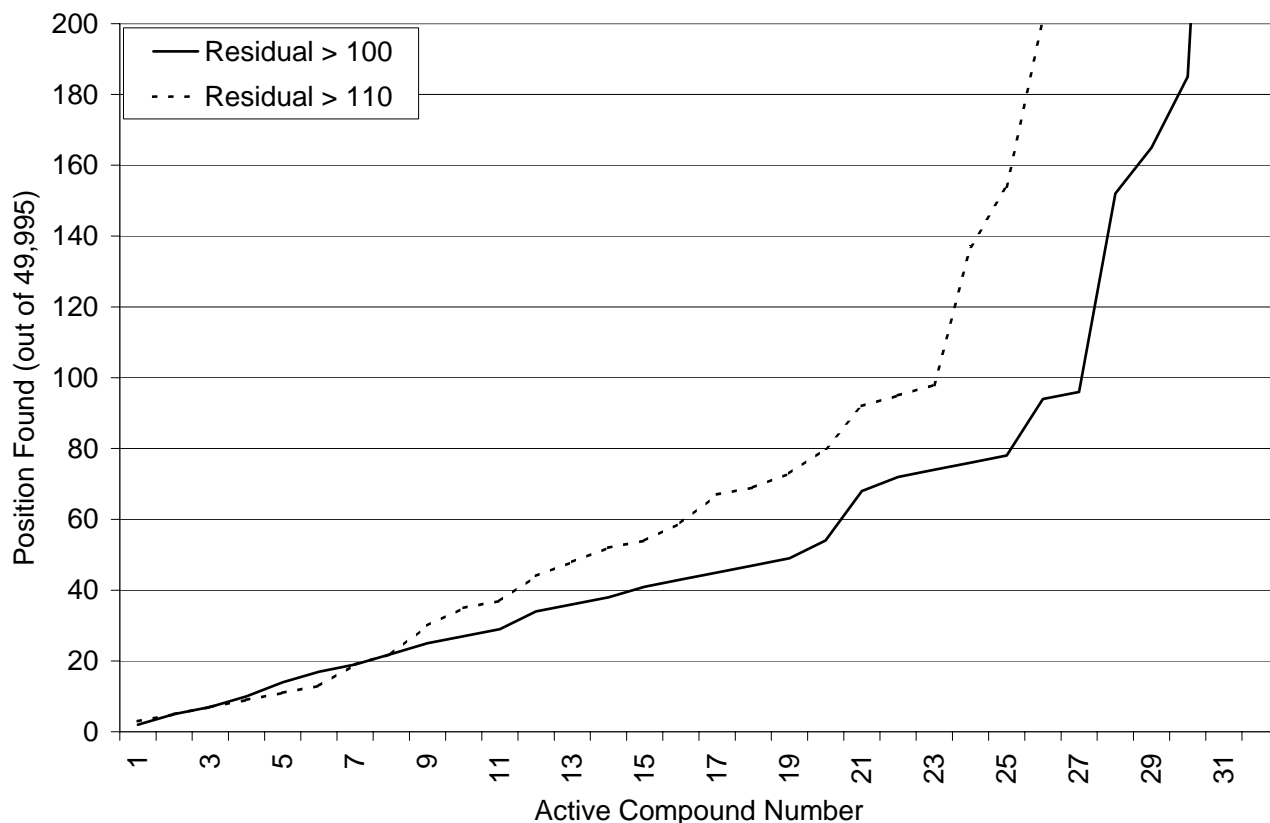


Figure 38. The first 30 active compounds (out of 47; 64%) are found in the top 200 out of 49,995 positions (0.4%) of the ranked training set. The definition of inactive structures as having average residual activity larger 100% performed slightly better than the alternative definition of larger 110%. The results on the test set employing this classifier were nonetheless sobering.

Sample structures predicted as being “active” are shown in Figure 39. Recurring features can readily be identified such as halogenated benzenes, pyrazoles, methylesters and 1,2,3 triazines.

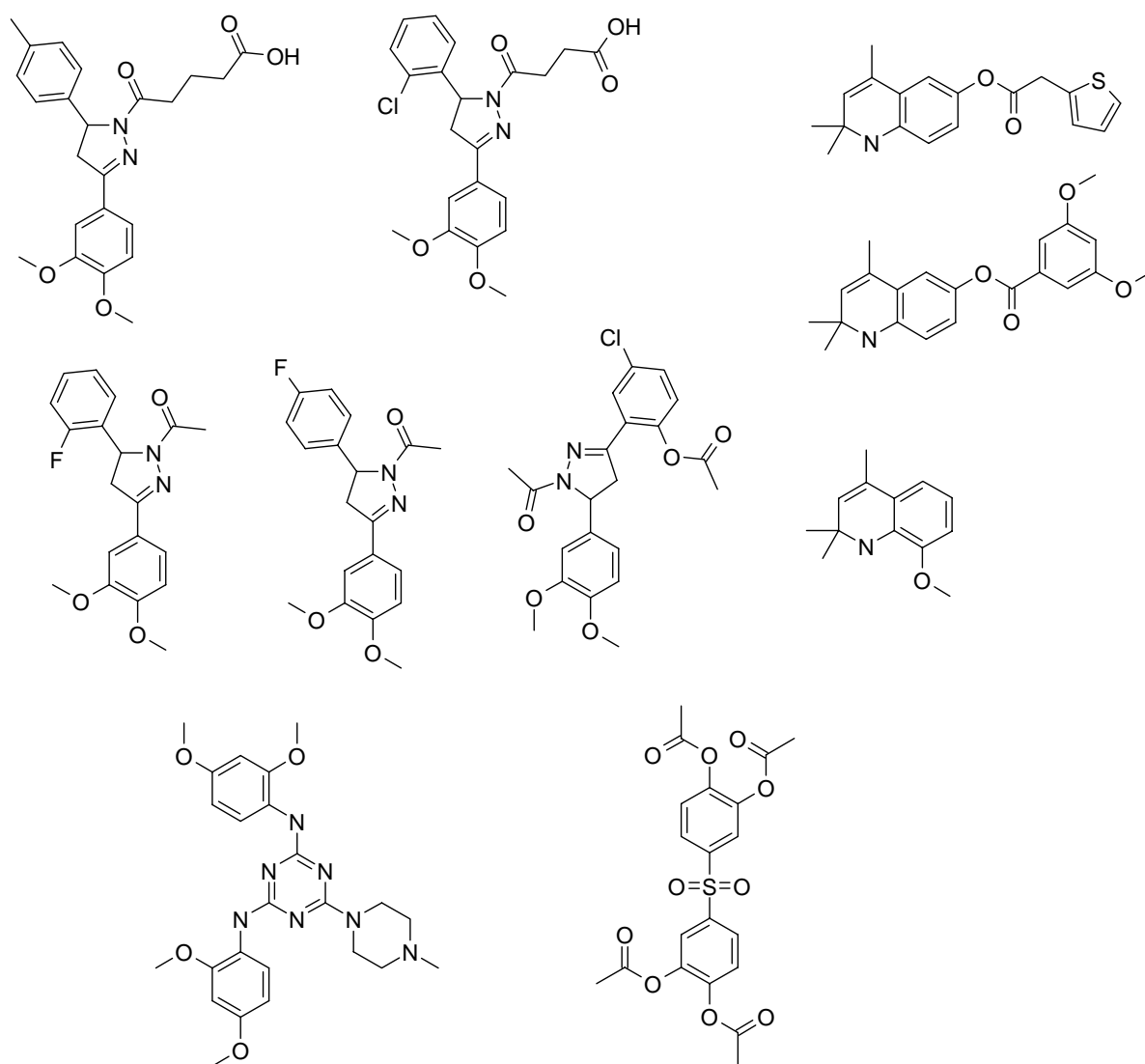


Figure 39. Structures initially predicted as being “active”, based on an active set defined by the 32 hits plus 15 additional known DHFR inhibitors. Recurring features can readily be identified such as halogenated benzenes, pyrazoles, methylesters and 1,2,3 triazines. Still, due to the different chemical constituency of training and test set none of those is active in practice.

Soberingly, upon receiving the HTS screening results of the test data set virtually no enrichment was found for the ranked list of compounds we submitted. As listed in the official results of the competition, 4 of 42 actives (less than 75% residual activity in both runs) were found in the top 2,500 positions, corresponding to an enrichment factor of only slightly less than 2 (nearly 10% of all actives are contained in 5% of the library). Nonetheless, this was still the best result achieved among groups who submitted a ranking of the full test set on one of the criteria applied (average

inhibition values). This is particularly interesting in the light that many of the submissions were also employing structure-based (docking) methods, which also were not able to improve performance. Therefore, the method presented here seems to present a good trade-off between computational expense and quality of the results – with the strong indication that further improvement of algorithms is necessary. Also, actives found in the test set were very weak and not competitive, indicating problems with the dataset employed as well. Our initial assumption that the cutoff thresholds for active and inactive datasets were inappropriate or that feature selection should be employed lead only to marginal improvement of the enrichment factor of the test set, up to about 3. This was achieved if the active compound set was defined slightly more loosely, by an average of less than 80% residual activity, compiling 76 structures in the active training set (data for this scenario are given in Table 28).

Table 28. Number of compounds identified (hit rates) and equivalent enrichment factors for the first 96, 384 and 1536 positions of the sorted library. While a maximum enrichment factor of around 3 can be achieved for the original training and test sets, a random 50/50 split of the merged library is able to achieve enrichments of around ten on the first 96-well plate, independent of the particular split used. This underlines the need for a similar chemical composition between training and test set.

First ... positions	Hit Rates			Enrichment Factors		
	96	384	1536	96	384	1536
Actives < 80% activity; Inactives > 100% activity, 200 Features	2	4	10	3.4	1.7	1.1
Ten-fold Random Validation Actives < 85% activity, Inactives > 100% activity, 200 Features	6.0 (0.7)	10.2 (2.4)	28.0 (3.0)	10.2 (1.2)	4.2 (1.0)	3.0 (0.3)

Examining training and test set more closely considerable differences in the chemical composition of both sets became apparent. This is illustrated in Figure 40, showing the five most potent structures from the training set and from the test set, respectively.

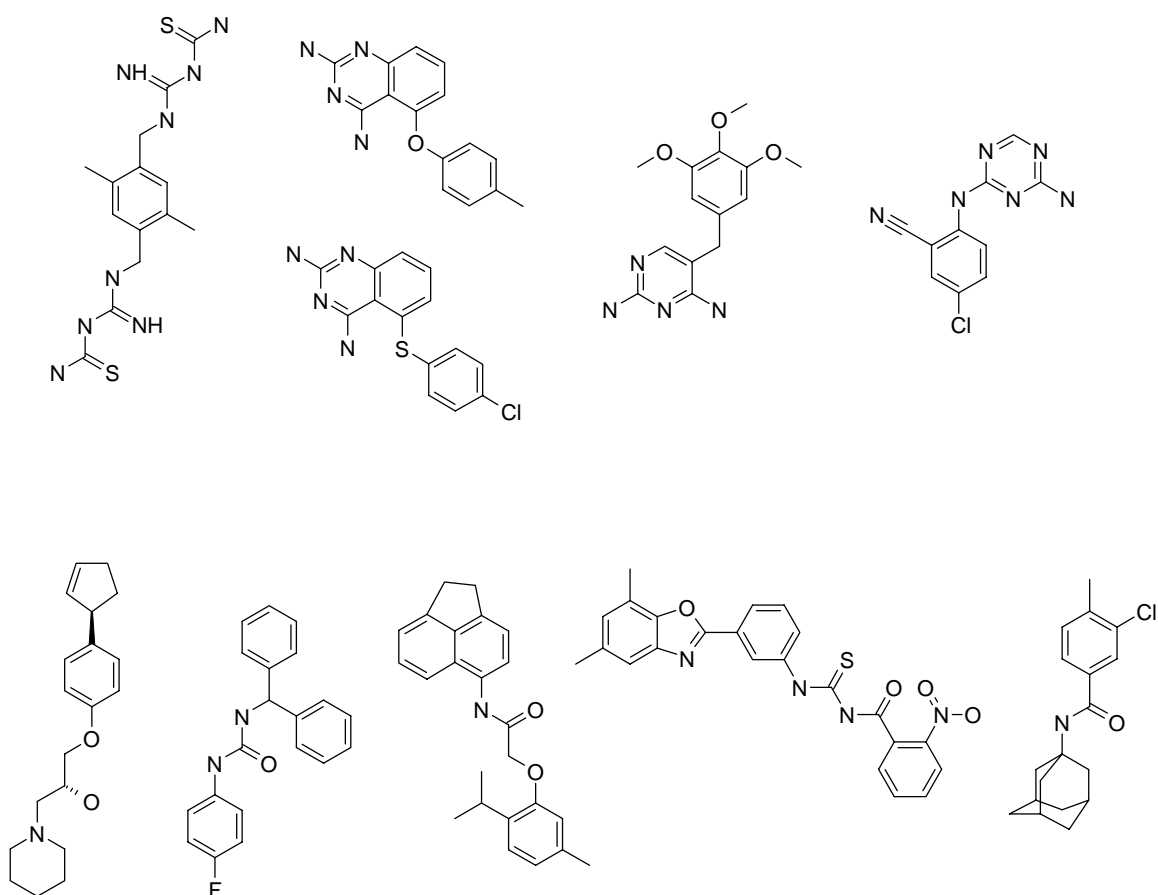


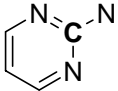
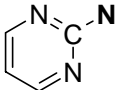
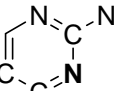
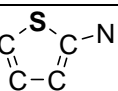
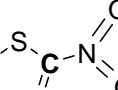
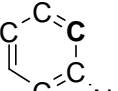
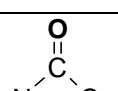
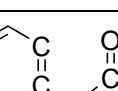
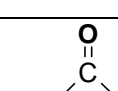
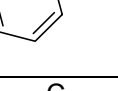
Figure 40. Comparison of the most potent inhibitors of the training set (upper half) and the most potent inhibitors of the test set (lower half). Structural differences can already be identified on this small subset of compounds, for example the high number of pyridazine rings and guanidin groups in the training set.

Fundamental structural differences can already be identified from this small subset of compounds, such as the high number of pyridazine rings and guanidin groups in the training set as opposed to the test set. From this small number of compounds also the “false-positive” predictions shown in Figure 39 can be explained, since all of the false-positives contain nitrogen-heterocycles, and all except one of them possess methoxy-moieties – just like the five most active compounds from the training set. In effect, in the first classification run a model has been built which was only able to distinguish inactive compounds of one sort from other inactive compounds in the test set, which also explains why feature selection was not able to improve performance. This anecdotal evidence is corroborated by a statistical analysis of fragments showing highest information gain in discriminating the active structures of training set and test set from each other, shown in Table 29. (Note that activity is here and in the following

defined as an average residual activity of less than 80% to increase the data basis.) The statistical analysis was carried out separately for the five fragments that showed highest information gain and were more frequent in the training set, and for the five fragments showing highest information gain which were more frequent in the test set. Those features characteristic for active structures in the training set, as opposed to active structures in the test set, are shown in the upper half of Table 29. The features occur much more often in the whole training set, compared to the whole test set (first two columns, ‘Number in Training Set’ vs. ‘Number in Test Set’), indicating a different chemical composition of both libraries. Even more profound is the relative frequency difference in active compounds from the training set and active compounds from the test set. While between eight and ten structures from the training set contain each of the activity-conferring features from the training set, *none* of the actives from the test set contains any of those features. This shows that in particular the chemical composition of *active* compounds from training and test set is different. Thus, it is not surprising that our initial enrichment factors were so low (this method is based solely on the molecular graph and hence requires detection of substructures that contribute to activity).

Features characteristic for active structures in the test set, as opposed to active structures in the training set, are shown in the lower half of Table 29. Again, those features conferring activity to the active structures of the test dataset are also much more frequent in the whole test dataset, compared to the whole training dataset. This corroborates the finding above that both data sets are overall of different chemical composition. It also applies – particularly – to the relative frequencies of activity-conferring features from the test set in the active parts of test and training set. While a high number of active compounds from the test set contains the characteristic features, this is rarely the case for the training set actives (for details see Table 29). This also means that the chemical composition of active compounds in the test set is different from the composition of the training set, with respect to the features shown in Table 29 but also for a much larger number of features which are not shown here.

Table 29. Features showing highest information-gain in discriminating active structures of the training set from those of the test set. Features characteristic for the most active compounds of each set are also more frequent in the whole set; this ratio is even more apparent among the active structures of each set.

Characteristic Features of Training-Set Active Structures				
	Number in Training Set	Number in Test Set	Among Actives in Training Set	Among Actives in Test Set
	416	159	10	0
	295	72	10	0
	40	72	8	0
	106	12	10	0
	136	0	9	0
Characteristic Features of Test-Set Active Structures				
	9077	18645	14	133
	6580	19186	4	126
	2449	10685	0	76
	9191	16043	12	115
	15202	25851	22	160

The discussion up to this point leads to two valid conclusions. On the one hand the work presented here shows that MOLPRINT 2D is not capable of finding completely novel hit structures on the data set given. This can probably be extended to related approaches that belong to the group of exact-fragment-matching similarity searching methods. They are not able to exploit knowledge about activity space from one chemical series and apply this knowledge to a different chemical series (in effect, to identify bioisosteres). This is not surprising due to the strict definition of features. On the other hand this result emphasizes the need for an even distribution of “chemistry” between the training and the test set when exact-fragment matching methods are employed. In order to examine the effect of a more equalized chemical composition between training and test set another calculation was performed.

Ten-fold random scrambling of the 99,995 compounds of both (training and test) datasets was performed in order to achieve comparable chemical compositions, resulting in new training and test sets of 49,995 compounds and 50,000 compounds, respectively. From the training set in each run, those compounds showing on average less than 85% residual activity (251 compounds) and those showing more than 100% residual activity (21,551 compounds) were used as “active” and “inactive” datasets, respectively. Thresholds for activity and inactivity were chosen to provide an optimum between certainty of activity and dataset size. It should be noted that in case of the new training and test sets performance did not vary greatly for definitions of active compounds below between 80% and 90% residual activity, and for definitions of inactive compounds above between 100% and 130% residual activity. 200 Features were selected in each of the ten cross-validation runs. Again, performance did not vary greatly between a selection of 200 features, up to a maximum number of features defined by the number of features present in the active dataset (e.g., 576 features for 76 active compounds at an 80% threshold).

In each of the ten runs an average of 6 active compounds were identified in the first 96 positions, an average of 10 active compounds in the first 384 positions and an average of 28 active compounds in the first 1536 positions. This corresponds to enrichment factors of around 10, 4 and 2, respectively, based on a total number of 307 actives (defined as average residual activity <80%). Sample structures identified as “actives” in one of the ten cross-validation runs are shown in Figure 41. Three different scaffolds are identified among the seven hits obtained. This illustrates both the capability of the method to identify actives, as well as the capability to identify

multiple scaffolds contributing to activity. Standard deviations of hit rates are small (around 1), showing robust performance of the method with respect to the particular composition of the training set used – provided that the chemistry between training and test set is comparable.

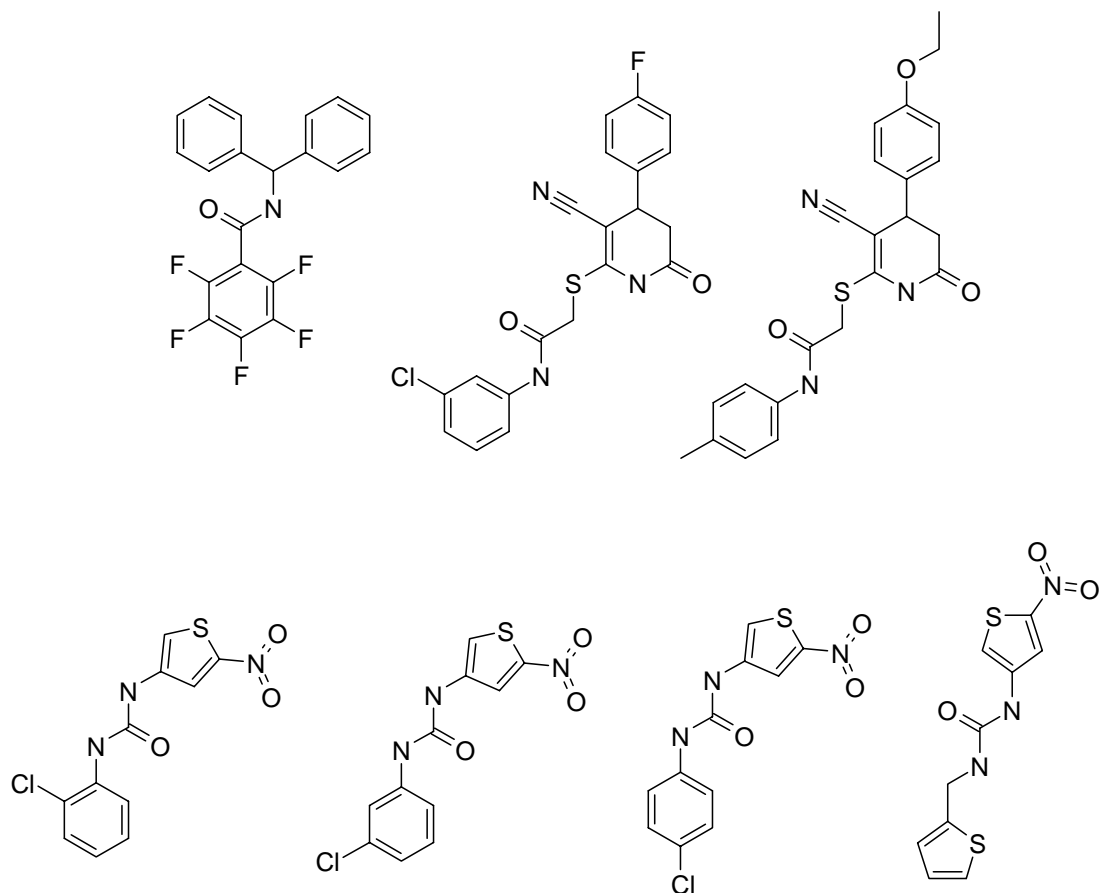


Figure 41. Active compounds identified in one of the random selection runs, discovering seven hits (average residual activity < 80%) on the first 96-well plate, equivalent to an enrichment factor of around ten. Three different scaffolds are identified in this case, illustrating the capability of discovering diverse molecules in the test set if it contains similar chemistry to the training set.

10.3 Conclusions

A fragment-based similarity searching method, MOLPRINT 2D, was employed for virtual screening of inhibitors of dihydrofolate reductase (DHFR) of *E. coli*. Performance on the original training and test sets was initially not satisfactory, with enrichment factors of no larger than three. This was found to be caused by a different chemical composition of training and test set. A similar chemical composition of training and test set lead to enrichment factors of around 10 in a ten-fold cross-validation study on the first 96-well plate.

The conclusions are two-fold: On the one hand the work presented here shows that MOLPRINT 2D is not capable of finding completely novel hit structures on the data set given. This can probably be extended to related approaches which belong to the group of exact-fragment-matching similarity searching methods. Still, they are able to combine knowledge from multiple active structures to give novel combinations of features, as shown previously. On the other hand this work emphasizes the need for an even distribution of “chemistry” between the training and the test set which can be implemented by careful experimental library designs.

11 On the information content of molecular descriptors with increasing level of sophistication

While in the previous chapter restrictions of circular descriptors are discussed, we will now challenge the information content of other types of descriptors. To do this, their information content is compared to very simple descriptors with increasing complexity levels.

Conventionally, so-called “enrichment factors” are calculated in the literature to establish a baseline for assessing the quality of virtual screening methods. Enrichment factors describe the number of active compounds found by employing a certain virtual screening strategy, as opposed to the number of compounds hypothetically found if compounds were screened randomly. Enrichment factors are defined by the success of the virtual screening algorithms at ordering the library, with the most likely compounds to be active being suggested by the algorithms to be screened first. Enrichment factors can range from 1, where molecules are sorted randomly and virtually no ‘enrichment’ is achieved by the algorithm, to >100 in which only a small percentage of the library needs to be screened to find a large number of active molecules^{270,271}. Very high enrichment can only be observed for very small parts of the sorted library due to the way the performance measure is calculated (see formula in material and methods section). Enrichments smaller than 1 are indicative of a preferred selection of inactive compounds, corresponding to an enrichment of active compounds at the bottom of the sorted library.

At first sight, this seems to be a suitable benchmark since it compares the “rational” drug discovery procedure to the “irrational”, or random method. As we demonstrate here enrichment factors, which are often reported in the double-digit numbers, overestimate the real performance of virtual screening algorithms by using too low a benchmark (random hit rates) for comparison. Enrichment factors of 10 create the impression that the particular virtual screening method saves a great amount of time, expenditure and experimental work. While indeed, at least in a retrospective sense, enrichment factors of that size are correct in the abstract way they are calculated, they are not a realistic benchmark for added value. The main reason for this probably lies in the fact that we assume that molecules in a given compound library are “random” (show every arrangement of molecular features with the same likelihood) – and hence

apply random selection methods to the library as a benchmark. However, there is no such thing as a “fully comprehensive random library” in which all possible arrangements of features are present in a uniform distribution, so that every library shows a bias with respect to one set of properties or another.

We show that enrichment factors of around four (averaged over several classes of active compounds and two different datasets) can be achieved using very simple, non-structural features. Atom counts distinguished by elements are used as a ‘dumb’ descriptor and the sum of absolute distances of the atom-count “fingerprint” is used as a similarity (or rather distance) measure. Enrichment factors obtained *via* this method on a dataset derived from the MDL Drug Data Report (MDDR) database²⁴⁷ are higher than e.g. those achieved *via* a virtual affinity fingerprint based method (DOCKSIM). This is particularly surprising in the view that docking based methods have a large amount of information about the binding site at hand (and are computationally very demanding).

On a different data set that is also derived from the MDDR, “virtual screening” using atom count vectors outperforms Unity fingerprints on some activity classes. While overall Unity fingerprints are still superior, their added value (enrichment rate with respect to these very simple features) drops down to a factor of around 2. This is particularly surprising since atom counts do not capture structural information at all, which is exactly the kind of information captured in more complex fingerprints.

On the other hand, other simple molecular features such as molecular weight are shown to give much lower enrichment. The simple assumption that it is possible to identify close active analogues by weight-based virtual screening can thus be dismissed. Interestingly, as shown below, count vectors of atoms as descriptors for molecules do not only retrieve a high number of active compounds, they are also able to identify actives from different structural classes.

Following the simplicity principle (simple models are more likely to produce the data observed, compared to more complex models) it can be concluded that the performance of virtual screening methods should be seen in relation to their complexity, and more sophisticated methods do not lead to superior results in every case. While this work on the one hand shows that the added value of virtual screening methods is not as large as might be expected from random-selection based enrichment factors, it also shows that there are fundamental differences present between different drug activity classes which can be detected by simple atom counts. These descriptors

(partially) implicitly capture overall properties such as size, lipophilicity, hydrogen bonding capabilities and polarity which seem to be suitable discriminants for some activity classes: in those cases they are as discriminating as structure-based fingerprints.

In the recent literature, several publications appeared which are related to the work presented here. In his review on one-dimensional descriptors²⁷², Livingstone discusses overall molecular parameters which are able to discriminate between compounds showing different physicochemical or biological behaviour. For example, blood-brain barrier penetration is closely related to logP, and electron density on a nitrogen atom in the HOMO of a set of aniline mustards and tumour inhibition can be related in a simple linear fashion. The focus is different in the present work where we apply simple molecular descriptors to ligand-based virtual screening, which involves multiple activity classes, in order to gauge how many active molecules can be found in retrospective virtual screening runs by their application. Results are compared to both random selection as well as structure-based descriptors, and this is performed in a quantitative fashion.

One of the simple descriptors employed by us is molecular weight, whose influence on target- (instead of ligand-) based virtual screening has been investigated by Pan *et al.*²⁷³. In this earlier work it was found that heavier molecules are favoured by docking algorithms due to the simple fact that on average more atom-atom interactions are present which contribute to the predicted binding energy. As a remedy normalization of the binding energy with respect to the number of heavy atoms per molecule (or a root thereof, depending on whether drug-like or lead-like compounds were wished) was suggested.

Bioactivity profiles (BPs) which include the number of hydrogen bond donors and acceptors, molecular weight, a kappa shape index and the number of rotatable bonds as well as the number of aromatic rings, were introduced by Gillet *et al.*²⁷⁴. BPs found application in distinguishing molecules from the World Drug Index and those from the SPRESI database (which were assumed to be inactive). Using single features such as the number of hydrogen bond donors alone enrichments of up to 4.6 were found in identifying WDI molecules in a merged dataset. While the simple description of the structures resembles our approach, Gillet *et al.* do not employ atom count vectors, which perform favourably in the work presented here. In addition we employ clearly defined targets instead of therapeutic classes in order to predict activity on a particular

receptor or enzyme, which is the usual aim in drug discovery programs. Similar differences in scope are given to the work by Wilton *et al.*²⁷⁵ who employed substructural analysis and bioactivity profiles in combination with Binary Kernel Discrimination (among other methods) for the identification of active compounds from the well-known NCI AIDS dataset. The activity-determining factor of the NCI AIDS dataset is growth inhibition, so neither a consistent molecular target nor multiple activity classes are employed.

Attempts to avoid “artificial enrichment”, defined as the identification of active compounds by differing simple molecular properties, were presented by Verdonk *et al.*¹⁶³ in the context of target-based virtual screening. Considering heavy atom counts alone on two hypothetical libraries of active compounds, which are either on average much heavier or much lighter than the whole library, was shown to give considerable enrichments. Several steps were taken in joining a library of true active binders and non-active HTS compounds to give a similar distribution of features in both datasets in order to eliminate “artificial enrichment”. In our case we expand the work of Verdonk *et al.*, whose main focus was on the performance of docking-based virtual screening methods, to multiple simple molecular features as well as multiple activity classes. Interestingly, the consideration of molecular weight alone did not give major improvements in identifying active compounds over random selection in our ligand-based virtual screening setting. This may simply depend on the fact that most of the activity classes are not as different in molecular weight from the distribution of the whole library as in the hypothetical example given by Verdonk and co-workers.

To summarize, the novelty of the work presented here is that it employs simple features for ligand-based virtual screening on a large dataset (>100k compounds), which due to its size and comprehensiveness can be seen as one of the best publicly available ones. The dataset comprises multiple (eleven) activity classes for clearly defined molecular targets. Using simple features on this dataset, considerable enrichments can be found which are sometimes as good as those obtained using structural fingerprints. We believe that the performance of more complex descriptors employed for similarity searching has thus to be gauged in relation to their sophistication.

11.1 Material and methods

Two different datasets were examined which were previously subject to retrospective virtual screening using a variety of methods. The first dataset was published by Briem and Lessel¹³⁰ and it contains 957 ligands extracted from the MDDR database which was also used previously in this work. The set contains 49 5HT3 Receptor antagonists (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 574 compounds were selected randomly which did not belong to any of these activity classes. The second and larger dataset was presented recently by Hert *et al.*^{31,211} and it was also used above. 11 sets of active structures were defined, ranging in size from 349 to 1236 structures. (Full details of the dataset sizes are given in Table 2.). This dataset spans a variety of targets as well as a very large number of compounds, which provides a useful benchmark for a similarity searching method. One has to be aware of the shortcomings of all current drug database-derived datasets, which is the occurrence of close analogues, which favours 2D methods, and the fact that the MDDR does not include explicit information about inactivity of compounds (what this means is that inactive compounds identified as false-positives may well be active and thus true-positives). Nonetheless relative values on retrospectively analyzed datasets can be used to judge relative performance of different molecular similarity searching methods.

For all compounds of both datasets, simple atom count vectors were calculated using MOE²⁵², namely the total number of atoms, the number of heavy atoms and the numbers of Boron, Bromine, Carbon, Chlorine, Fluorine, Iodine, Nitrogen, Oxygen, Phosphorus and Sulphur atoms. Thus no structural descriptors at all were contained in this “fingerprint” representation which, besides the compound ID, contains just 12 integer numbers describing the frequency of different elements in the molecule.

Single queries were selected from both of the datasets. On the first, smaller dataset each “active” compound was selected once and the remaining structures were sorted according to their similarity to the query. Hit rates among the ten nearest neighbours were reported, following the earlier protocol by Briem and Lessel¹³⁰ which compared Feature Trees⁷¹, ISIS MOLSKESYS²³⁹, Daylight Fingerprints⁸⁶, SYBYL Hologram QSAR Fingerprints⁸⁷ and FLEXSIM-X¹³⁰, FLEXSIM-S²⁴⁰ and DOCKSIM¹²⁹ virtual

affinity fingerprints. From the second, larger dataset, single queries were selected randomly ten times and the number of active compounds in the top 5% of the sorted library was recorded, allowing comparison to Unity fingerprints³¹ and circular fingerprints (MOLPRINT 2D)^{79,80}.

For both datasets, the fraction / number of active compounds found was compared between the screening rungs using “dumb” atom count vector descriptors and (supposedly) more information-rich 2D and 3D descriptors. Enrichment factors (E_f) after x% of the focused library were calculated according to the following formula ($N_{\text{experimental}}$ = number of experimentally found active structures in the top x% of the sorted database, N_{expected} = number of expected active structures, N_{active} = total number of active structures in database, N_{total} = total number of structures in database).

$$E_f = \frac{N_{\text{experimental}}^{x\%}}{N_{\text{expected}}^{x\%}} = \frac{N_{\text{experimental}}^{x\%}}{\left(\frac{N_{\text{active}}}{N_{\text{total}}} \cdot x\%\right)}$$

11.2 Results and discussion

The average hit rate using “dumb” atom count-descriptors, compared to a variety of 2D and 3D similarity searching methods is shown in Figure 42 for the first, smaller dataset.

Atom count descriptors achieve an enrichment of about 4-fold (average hit rate of 34%), compared to a hit rate from random sampling of 7.9%. Hit rates vary through the five datasets of active compounds between 16% (Thromboxane A2 Antagonists) and 51% (HMG-CoA Reductase Inhibitors). 2D based fingerprints (MOLPRINT 2D, Feature Trees, ISIS MOLSKEYS, Daylight Fingerprints, SYBYL Hologram QSAR Fingerprints; results taken from Briem and Lessel¹³² and Bender *et al.*⁸⁰) are found overall to be superior to virtual affinity fingerprints (FLEXSIM-X, FLEXSIM-S and DOCKSIM). Elemental atom counts achieve an average hit rate of 34% and are thus ranked worse than FLEXSIM-X and FLEXSIM-S, but better than DOCKSIM. On this first dataset, computationally much more demanding virtual affinity fingerprints (the DOCKSIM method) are outperformed by simple atom counts. While docking-based methods in principle are able to exploit a wealth of information of the binding site, this confirms – in agreement with other recent research^{276,277} – that current scoring functions are not able to predict binding affinity of a ligand to a receptor reliably. 2D

descriptors are generally able to add value to similarity searching protocols, but not as much as a random screening baseline suggests.

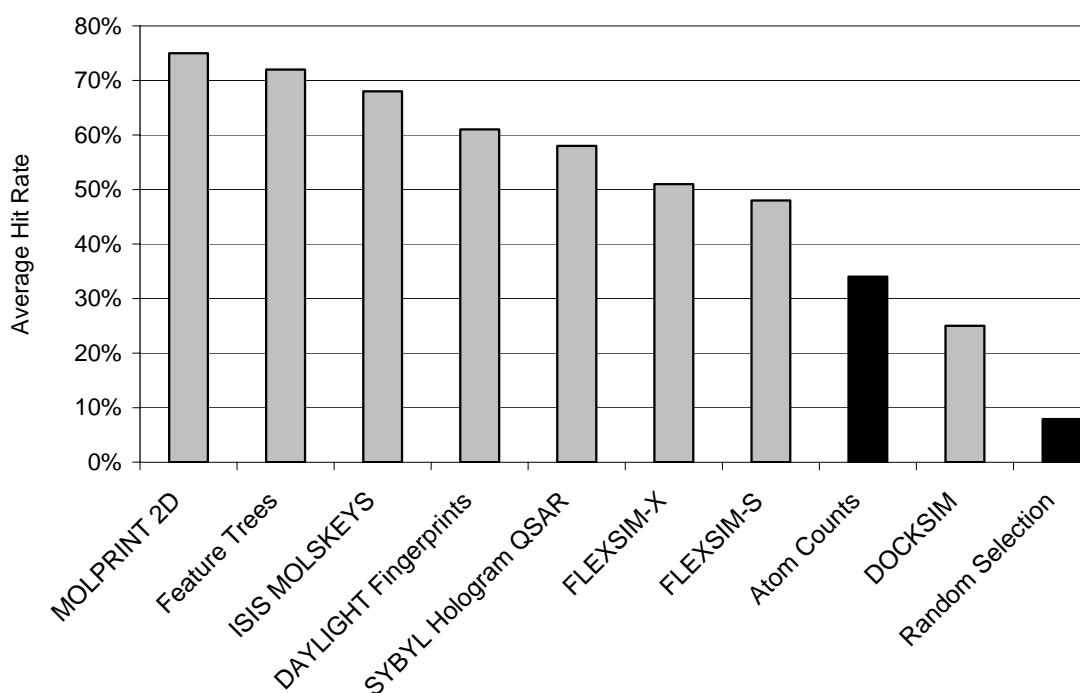


Figure 42. The average hit rate using “dumb” atom count descriptors, compared to a variety of 2D and 3D similarity searching methods. Even atom count descriptors achieve an enrichment of about 4-fold which is already superior to one of the virtual affinity fingerprint methods, DOCKSIM and around half the enrichment achieved by other methods employed.

The average fraction of active compounds retrieved on the second, large dataset are given in Table 30 and Figure 43. Between 1.7-fold and 13-fold enrichment can be observed for simple atom counts within the top 5% of the sorted database. Lowest enrichment was observed for Cyclooxygenase Inhibitors (enrichment of 1.7) and 5HT Reuptake Inhibitors (enrichment of 2.3), highest enrichment for Renin Inhibitors (13.3-fold enrichment) and HIV Protease Inhibitors (enrichment of 5.6). Compared to virtual screening results using Unity fingerprints, simple atom counts are able to outperform Unity fingerprints in two instances, namely on the 5HT3 and Dopamine D2 Antagonist datasets, while in other cases Unity fingerprints are superior to simple atom counts by a factor of up to 2. While circular fingerprints (MOLPRINT 2D) retrieve overall more active compounds, in cases such as the 5HT3 Antagonist dataset simple atom counts are not outperformed by a large margin. All three methods retrieve a remarkably similar number of compounds.

Hit Rates via Atom Counts, Unity Fingerprints and MOLPRINT 2D

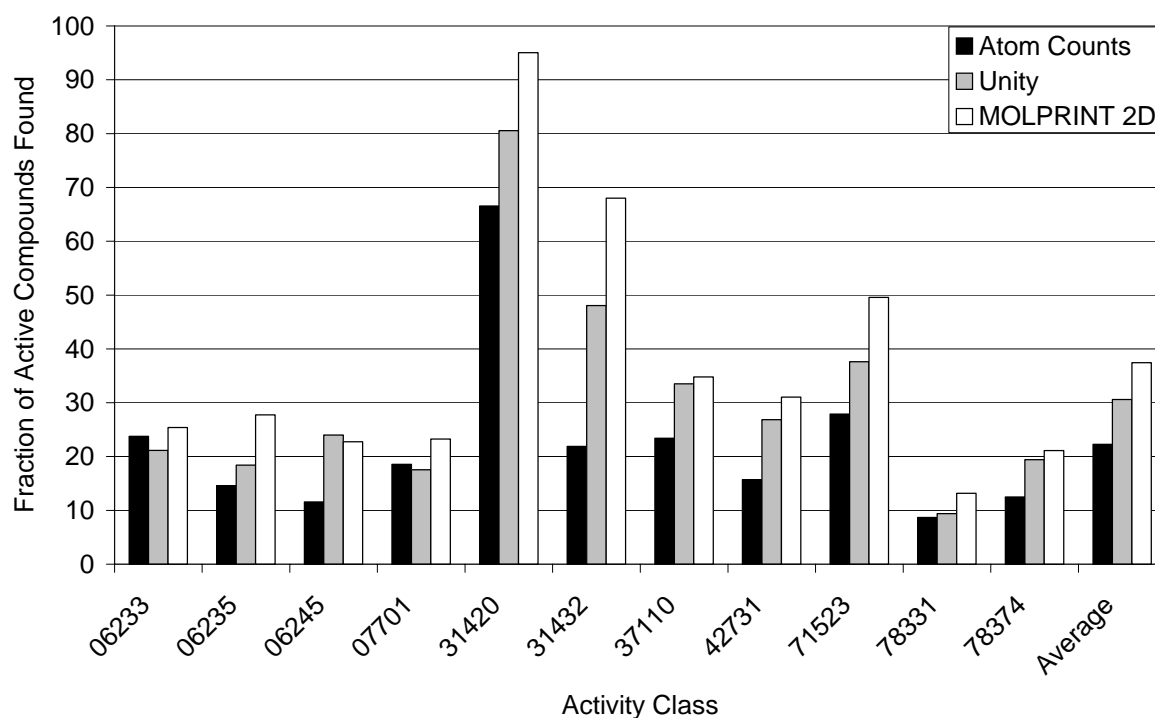


Figure 43. Fraction of active compounds found using simple atom counts, in comparison to Unity fingerprints. While Unity fingerprints outperform atom counts overall this margin is smaller than one might expect, given the fact that atom counts do not contain any structural information whatsoever while Unity fingerprints have that information available.

Table 30. Activity class, hit rate among the top5% of the sorted database and hypothetical enrichment for the different sets of active compounds of the large test set. Using simple atom count descriptors, up to more than ten-fold enrichment can be observed which is close to results achieved using Unity fingerprints on the same dataset.

Activity Class	06233	06235	06245	07701	31420	31432	37110	42731	71523	78331	78374	Average
Hit Rate Atom Counts	23.78	14.59	11.59	18.58	66.53	21.89	23.42	15.69	27.89	8.69	12.48	22.28
Enrichment	4.76	2.92	2.32	3.72	13.31	4.38	4.68	3.14	5.58	1.74	2.50	4.46
Hit Rate Unity	21.15	18.43	24.02	17.53	80.54	48.04	33.51	26.87	37.60	9.39	19.42	30.59
Unity / Atom Counts	0.89	1.26	2.07	0.94	1.21	2.19	1.43	1.71	1.35	1.08	1.56	1.43
Hit Rate MOLPRINT 2D	25.40	27.73	22.75	23.24	95.04	68.01	34.79	31.03	49.56	13.16	21.13	37.44
MOLPRINT 2D / Atom Counts	1.07	1.90	1.96	1.25	1.43	3.11	1.49	1.98	1.78	1.51	1.69	1.68

On this second dataset, Unity fingerprints are superior to simple atom counts only on some of the activity classes. Put another way, Unity fingerprints do not capture more information relevant to activity on some datasets than simple atom counts. Indeed they perform marginally worse than atom counts in those cases. If the performance of Unity fingerprints is superior, their added value amounts to a factor between 1 and 2, with circular fingerprints performing only slightly better. This result is particularly remarkable since non-structural information is commonly not expected to give nearly as good results for screening as when utilizing structural information. While Unity fingerprints were chosen in this work as a reference method it should be noted that they were employed simply because of their availability and ubiquity, not because of their particularly good or bad performance.

We also employed other simple molecular descriptors on this dataset to investigate whether the good result of atom count vectors is simply due to differences in size which is already known to have a profound impact on target-based (instead of ligand-based) virtual screening (docking). It is well-known²⁷³ that docking favours larger molecules simply due to the larger number of interactions present between ligand and target. In Figure 44 enrichment factors for the first 5% of the sorted library are given for MOLPRINT 2D structural fingerprints, Unity fingerprints, atom count vectors, the total number of atoms and molecular weight for the eleven classes of active compounds. Enrichment factors using molecular weight and the number of atoms alone show similar performance in that they only give meaningful enrichment in case of two datasets, Renin inhibitors and HIV Protease inhibitors, which are on average much heavier than the rest of the database. For other compound classes molecular weight and the total number of atoms are not meaningful discriminants between activity classes. Thus, the number of active compounds found by atom count vectors cannot be reduced to their ability to capture differences in size and molecular weight.

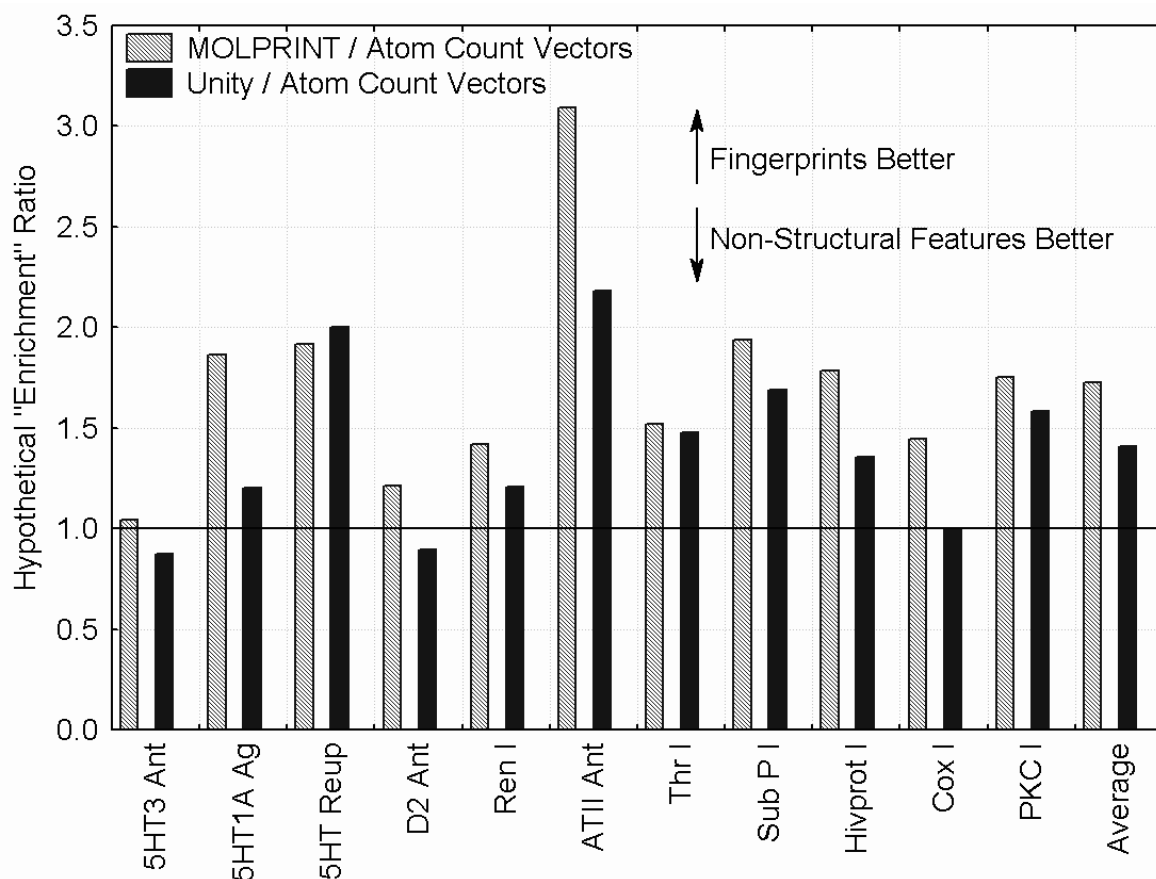


Figure 44. Enrichment factors obtained within the top 5% of the sorted library, depending on the activity class and for the descriptors MOLPRINT 2D and atom count vectors. Atom count vectors perform surprisingly well with average enrichments of >4 and they give in two cases (namely the D2 antagonist and the Cox inhibitor activity classes) results comparable to Unity fingerprints.

The overlap of active compounds retrieved using different descriptors is shown in Table 31, together with the highest theoretically achievable overlap given in brackets (which is smaller than 100% due to the different hit list sizes). Overall, overlap between 21% and 54% between the hit lists can be observed. Interestingly, hit lists from circular fingerprints and atom count vectors overlap by less than 50% of the theoretically possible value, indicating partly orthogonal behaviour of those descriptors.

Table 31. Overlap of the active compound set identified by employing structural features (circular fingerprints, MOLPRINT 2D), atom count vectors, the total number of atoms and the molecular weight (highest possible overlap if smaller set is completely contained in larger set in brackets). Overall medium overlap can be observed; with the two best performing methods (circular fingerprints and count vectors) showing less than 50% overlap of the retrieved active compounds.

	MOLPRINT 2D	Count Vector	#Atoms	Mol. Weight
MOLPRINT 2D	100%			
Count Vector	35% (81%)	100%		
#Atoms	25% (40%)	54% (57%)	100%	
Mol. Weight	21% (35%)	35% (55%)	38% (88%)	100%

To investigate this route further, retrieved active compounds were inspected manually, an example of which is given in Figure 45. Shown is the 5HT3 Antagonist used to perform virtual screening on the database at the top of the figure as well as two hit lists of active compounds, depicted below. On the left hand side active compounds are shown which are retrieved using circular fingerprints and the Tanimoto coefficient in the first 20 positions of the sorted list. On the right hand side active compounds are given which are retrieved using atom count vectors within the first 50 positions of the sorted list. While the absolute number of active compounds is larger in case of structural fingerprints, it is interesting to see that atom count vectors retrieve not only close analogue compounds. Instead a variety of scaffolds are found using this approach. This contradicts the simple assumption that atom count vectors are only able to identify close analogues, which, given their similar overall structure, might also be assumed to show similar atom count vectors.

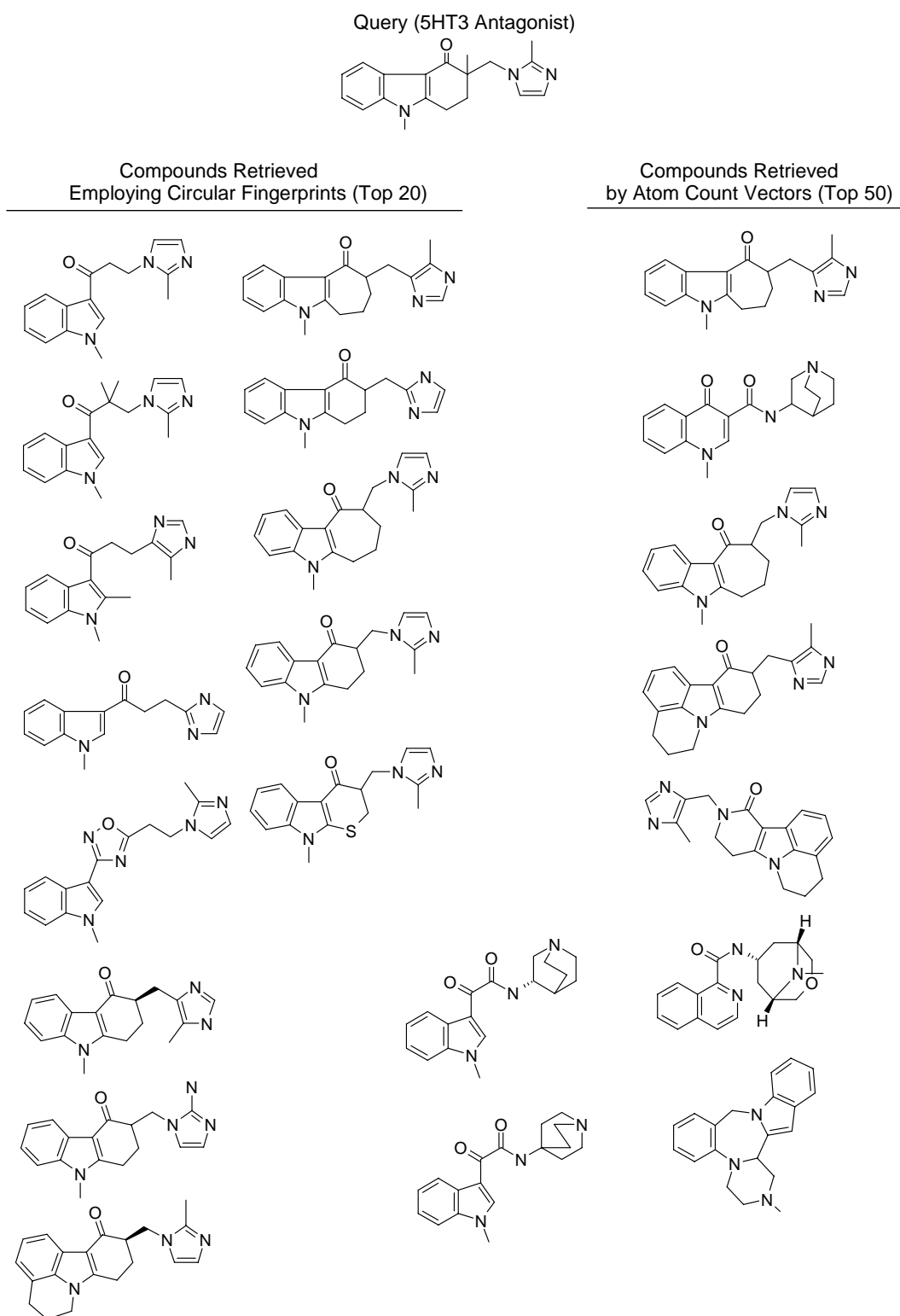


Figure 45. Compounds retrieved using the query (5HT3 Antagonist) at the top of the page and employing structural information (circular fingerprints; hits on the left) and simple atom counts (structures on the right). Although the number of active compounds found is larger in the case when structural fingerprints are employed enrichment is considerable by using atom count vectors alone. Interestingly, not only close analogues of the query compounds are identified by the atom count “descriptor”.

11.3 Conclusions

“Dumb” molecular features such as atom count vectors, which do not contain any structural information, are able to achieve “enrichment factors” of around 4 in ligand-based virtual screening and in some cases they even outperform Unity fingerprints. This puts previously reported double-digit figures for “enrichments” into perspective, showing that also simple methods are able to show considerable selection abilities on the datasets examined here. It follows that performance measures reported for more sophisticated virtual screening methods should be seen in relation to their complexity. Based on these results, it would seem that virtual screening methods do not add as much value as might be inferred from comparisons to hit rates achieved from random screening. On the other hand, simple atom counts seem to capture a lot of information about the difference between structures from different activity classes, implicitly encoding some of the global molecular parameters. This is seen only in two particular cases that we examined and to a much lesser extent true for molecular weight and the total number of atoms, which are known to be important determinants of predicted activity in target-based virtual screenings. It follows that there may be physicochemical properties important for binding which are (partly) implicitly captured by atom count vectors, such as size, hydrogen bond capabilities and polarity of the molecule, which are not contained in the total number of atoms or the molecular weight.

Put another way, the information content of two common structure-based descriptors for virtual screening purposes is in some cases not higher than the non-structural information about the number of atoms per element in the structure. The extent of this finding depends on the class of active structures. At the same time overlap between compounds retrieved based on structural features (circular fingerprints) and atom count vectors is rather low, suggesting some orthogonality in the features they describe. While “enrichment factors” obtained by atom count vectors are usually lower than those obtained by using structural descriptors, retrieved active compounds show a surprising variety of scaffolds as exemplified by a 5HT3 Antagonist virtual screening run, contradicting the assumption that close analogues are mainly identified by atom count vectors.

As a bottom line, the limitations of the ‘molecular similarity principle’ in the context of virtual screening should never be forgotten: small structural changes may, in the

arena of bioactivity, give rise to large changes in activity space. This may partly explain the problem in finding suitable molecular descriptors for this task and thus also the relatively good performance of very simple approaches to the problem.

12 Epilogue

My personal motivation of the present work was the development of a “groundbreaking” new approach to molecular similarity searching. I admit that this did not happen. Still, I learned a lot about current approaches to the problem and, probably even more important, about the restrictions we have to live with in this area. While similarity searching works to a certain extent and it is computationally cheap, it still shows significant shortcomings, resulting partly from the way it is validated (on only partially suitable datasets) and partly from the underlying assumptions. While the author sincerely hopes that one day a description of structures is devised that ‘understands’ which structural changes influence changes in biological response, it is also acknowledged that this goal is far from being achieved today.

References

Formatted according to *Journal of Medicinal Chemistry* guidelines.

- (1) Itskowitz, P.; Tropsha, A. Nearest neighbors QSAR modeling as a variational problem: Theory and applications. *J. Chem. Inf. Model.* **2005**, *45*, 777-785.
- (2) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3-50.
- (3) Dill, K. A. Additivity principles in biochemistry. *J. Biol. Chem.* **1997**, *272*, 701-704.
- (4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
- (5) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
- (6) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049-3059.
- (7) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350-4358.
- (8) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204-3218.
- (9) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103-108.
- (10) Sheridan, R. P. Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037-1050.
- (11) Dean, P. M. *Molecular Similarity in Drug Design*; Kluwer Academic Publishers: Dordrecht, 1994.
- (12) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2002**, *7*, 903-911.
- (13) Bohm, H.-J.; Schneider, G. *Virtual Screening for Bioactive Molecules*; Wiley-VCH, Weinheim, 2000.
- (14) Schneider, G.; Bohm, H. J. Virtual screening and fast automated docking methods. *Drug Discov. Today* **2002**, *7*, 64-70.
- (15) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295-307.
- (16) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003**, *22*, 151-185.
- (17) Directive 2003/15/EC of the European Parliament.
- (18) Greene, N. Computer systems for the prediction of toxicity: an update. *Adv. Drug Deliv. Rev.* **2002**, *54*, 417-431.
- (19) Thomson ISI Web of Knowledge - <http://www.isinet.com>.
- (20) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: Dordrecht, 2003.
- (21) Gasteiger, J. *Handbook of Chemoinformatics*; Wiley-VCH: Weinheim, 2003.
- (22) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening - an overview. *Drug Discov. Today* **1998**, *3*, 160-178.
- (23) Gillet, V. J.; Wild, D. J.; Willett, P.; Bradshaw, J. Similarity and dissimilarity methods for processing chemical structure databases. *Comput. J.* **1998**, *41*, 547-558.
- (24) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233-245.
- (25) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity - A review. *QSAR Comb. Sci.* **2004**, *22*, 1006-1026.
- (26) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882-894.
- (27) Johnson, M.; Basak, S.; Maggiora, G. A Characterization of Molecular Similarity Methods for Property Prediction. *Math. Comput. Model.* **1988**, *11*, 630-634.
- (28) Kubinyi, H. Chemical similarity and biological activities. *J. Braz. Chem. Soc.* **2002**, *13*, 717-726.

- (29) Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327-354.
- (30) Goldstone, R. L.; MIT Press: Cambridge, MA, 2004.
- (31) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177-1185.
- (32) 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.
- (33) Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quant. Chem.* **1980**, *17*, 1185-1189.
- (34) Carbo-Dorca, R.; Besalu, E. A general survey of molecular quantum similarity. *THEOCHEM - J. Mol. Struct.* **1998**, *451*, 11-23.
- (35) 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.
- (36) Stiefl, N.; Baumann, K. Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure-activity relationship technique. *J. Med. Chem.* **2003**, *46*, 1390-1407.
- (37) Cramer, R. D. Topomer CoMFA: a design methodology for rapid lead optimization. *J. Med. Chem.* **2003**, *46*, 374-388.
- (38) Brugger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of descriptors from molecular structures. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 105-110.
- (39) Glen, R. C.; Rose, V. S. Computer-Program Suite for the Calculation, Storage and Manipulation of Molecular Property and Activity Descriptors. *J. Mol. Graph.* **1987**, *5*, 79-86.
- (40) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
- (41) Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547-579.
- (42) Adamson, G. W.; Bush, J. A. A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55-58.
- (43) 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. *Comb. Chem. High-Throughput Screen.* **2002**, *5*, 155-166.
- (44) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379-386.
- (45) 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.
- (46) Martin, Y. C. Euro QSAR 2002: Designing Drugs and Crop Protectants: Processes, Problems and Solutions, 2002.
- (47) Horvath, D.; Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680-690.
- (48) Godden, J. W.; Bajorath, J. An information-theoretic approach to descriptor selection for database profiling and QSAR modeling. *QSAR Comb. Sci.* **2003**, *22*, 487-497.
- (49) 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.
- (50) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325-3329.
- (51) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882-1889.
- (52) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3-25.
- (53) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235-249.

- (54) Dixon, S. L.; Merz, K. M., Jr. One-dimensional molecular representations and similarity calculations: methodology and validation. *J. Med. Chem.* **2001**, *44*, 3795-3809.
- (55) Wang, N.; Delisle, R. K.; Diller, D. J. Fast small molecule similarity searching with multiple alignment profiles of molecules represented in one-dimension. *J. Med. Chem.* **2005**, *48*, 6980-6990.
- (56) Balaban, A. T.; Balaban, T. S. Correlations Using Topological Indexes Based on Real Graph Invariants. In *J. Chim. Phys.-Chim. Biol.*, 1992; pp 1735-1745.
- (57) Balaban, A. T. Local Versus Global (Ie Atomic Versus Molecular) Numerical Modeling of Molecular Graphs. *J. Chem. Inf. Comput. Sci.*, 1994; pp 398-402.
- (58) Estrada, E.; Uriarte, E. Recent advances on the role of topological indices in drug discovery research. *Curr. Med. Chem.* **2001**, *8*, 1573-1588.
- (59) Wilkins, C. L.; Randic, M. A Graph Theoretical Approach to Structure-Property and Structure-Activity Correlations. *Theor. Chim. Acta* **1980**, *58*, 45-68.
- (60) Randic, M.; Wilkins, C. L. Graph theoretical approach to recognition of structural similarity in molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31-37.
- (61) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys.* **1982**, *89*, 399-404.
- (62) Balaban, A. T. Chemical Graphs - Looking Back and Glimpsing Ahead. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 339-350.
- (63) Hall, L. H.; Kier, L. B. Electrotological State Indexes for Atom Types - a Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039-1045.
- (64) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513-520.
- (65) 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*, 7668-7767.
- (66) Barnard, J. M. Substructure Searching Methods - Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532-538.
- (67) Free, S. M., Jr.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *53*, 395-399.
- (68) Cramer, R. D., 3rd; Redl, G.; Berkoff, C. E. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* **1974**, *17*, 533-535.
- (69) Rucker, G.; Rucker, C. On finding nonisomorphic connected subgraphs and distinct molecular substructures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 314-320.
- (70) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for 3-Dimensional Structure- Directed Quantitative Structure-Activity-Relationships .1. Partition-Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565-577.
- (71) 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.
- (72) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic Identification of Molecular Similarity Using Reduced- Graph Representation of Chemical-Structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639-643.
- (73) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338-345.
- (74) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The reduced graph descriptor in virtual screening and data- driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145-2156.
- (75) 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.
- (76) Faulon, J. L.; Visco, D. P., Jr.; 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.
- (77) Xing, L.; Glen, R. C. Novel methods for the prediction of logP, pK(a), and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796-805.
- (78) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK(a) by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870-879.
- (79) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170-178.

- (80) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors: evaluation of performance. In *J. Chem. Inf. Comput. Sci.*, 2004; pp 1708-1718.
- (81) 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.
- (82) 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.
- (83) Xue, L.; 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.
- (84) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1-9.
- (85) MDL Drug Data Report; MDL ISIS/HOST software, MDL Information Systems, Inc.
- (86) DAYLIGHT, Version 4.62, DAYLIGHT Inc., Mission Viejo, California, USA.
- (87) SYBYL, Tripos Inc., St. Louis, Minnesota, USA.
- (88) Burden, F. R.; Winkler, D. A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999**, *42*, 3183-3187.
- (89) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discov. Design* **1998**, *9-11*, 339-353.
- (90) Jones, G.; Willett, P.; Glen, R. C. *Pharmacophore perception, development and use in drug design*; International University Line, 2000; pp 48.
- (91) Kotani, T.; Higashiura, K. Rapid evaluation of molecular shape similarity index using pairwise calculation of the nearest atomic distances. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 58-63.
- (92) Meyer, A. Y.; Richards, W. G. Similarity of molecular shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 427-439.
- (93) Good, A. C.; Richards, W. G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112-116.
- (94) Andrews, K. M.; Cramer, R. D. Toward general methods of targeted library design: topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* **2000**, *43*, 1723-1740.
- (95) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric-Field. *Int. J. Quantum Chem.* **1987**, 105-110.
- (96) Walker, P. D.; Arteca, G. A.; Mezey, P. G. A Complete Shape Characterization for Molecular Charge- Densities Represented by Gaussian-Type Functions. *J. Comput. Chem.* **1991**, *12*, 220-230.
- (97) 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.
- (98) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99*, 3503-3510.
- (99) Bultinck, P.; Kuppens, T.; Girones, X.; Carbo-Dorca, R. Quantum similarity superposition algorithm (QSSA): a consistent scheme for molecular alignment and molecular similarity based on quantum chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1143-1150.
- (100) Besalu, E.; Girones, X.; Amat, L.; Carbo-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Accounts Chem. Res.* **2002**, *35*, 289-295.
- (101) Boon, G.; Langenaeker, W.; De Proft, F.; De Winter, H.; Tollenaere, J. P.; Geerlings, P. Systematic study of the quality of various quantum similarity descriptors. Use of the autocorrelation function and principal component analysis. *J. Phys. Chem. A* **2001**, *105*, 8805-8814.
- (102) Carbo, R.; Calabuig, B. Quantum Similarity Measures, Molecular Cloud Description, and Structure Properties Relationships. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 600-606.
- (103) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.
- (104) Wold, H. *Encyclopedia of Statistical Sciences*; Wiley: New York, 1985; pp 581.
- (105) 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.

- (106) Klebe, G. Comparative molecular similarity indices analysis: CoMSIA. *Perspect. Drug Discov. Design* **1998**, *12*, 87-104.
- (107) Marshall, C. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. The Conformational Parameter in Drug Design: The Active Analog Approach. *Computer Assisted Drug Design*; Am Chem Soc, 1979; pp 205-226.
- (108) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem.-Int. Edit.* **1999**, *38*, 2894-2896.
- (109) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128-136.
- (110) Gund, P. Three-dimensional pharmacophoric pattern searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117-143.
- (111) Bemis, G. W.; Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607-628.
- (112) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New method for rapid characterization of molecular shapes: applications in drug design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79-85.
- (113) 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.
- (114) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567-597.
- (115) Martin, Y. C. 3D database searching in drug design. *J. Med. Chem.* **1992**, *35*, 2145-2154.
- (116) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251-3264.
- (117) 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*, 1367-1387.
- (118) Mason, J. S.; Cheney, D. L. Ligand-Receptor 3-D Similarity Studies Using Multiple 4-Point Pharmacophores. *Pac. Symp. Biocomput.*, 1999; pp 456-467.
- (119) Mason, J. S.; Cheney, D. L. Library Design and Virtual Screening Using Multiple 4-Point Pharmacophore Fingerprints. *Pac. Symp. Biocomput.*, 2000; pp 573-584.
- (120) Clark, T. QSAR and QSPR based solely on surface properties? *J. Mol. Graph.* **2004**, *22*, 519-525.
- (121) Gaillard, P.; Carrupt, P. A.; Testa, B.; Boudon, A. Molecular lipophilicity potential, a tool in 3D QSAR: method and applications. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 83-96.
- (122) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface-Area Structural Descriptors in Computer-Assisted Quantitative Structure Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323-2329.
- (123) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315-2327.
- (124) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. IcePick: a flexible surface-based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60-66.
- (125) Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 199-213.
- (126) Ghuloum, A. M.; Sage, C. R.; Jain, A. N. 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.
- (127) Ihlenfeldt, W. D.; Gasteiger, J. Hash Codes for the Identification and Classification of Molecular-Structure Elements. *J. Comput. Chem.* **1994**, *15*, 793-813.
- (128) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107-118.
- (129) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401-3408.
- (130) 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.
- (131) Dixon, S. L.; Villar, H. O. Bioactive diversity and screening library selection via affinity fingerprinting. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192-1203.

- (132) Briem, H.; Lessel, U. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discov. Des.* **2000**, *20*, 231-244.
- (133) Soltzberg, L. J.; Wilkins, C. L. Molecular Transforms: A Potential Tool for Structure-Activity Studies. *J. Am. Chem. Soc.* **1977**, *99*, 439-443.
- (134) Schuur, J. H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334-344.
- (135) 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.
- (136) 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.
- (137) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discov. Des.* **1998**, *3*, 199-213.
- (138) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT). *IEEE Transact. Syst. Man Cybern.* **2004**, *5*, 4553 - 4558.
- (139) Istvan, E. S. Structural mechanism for statin inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *Am. Heart J.* **2002**, *144*, S27-32.
- (140) Almond - Molecular Discovery Ltd. (<http://www.moldiscovery.com>).
- (141) Kastenholtz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E.; Fox, T. GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.* **2000**, *43*, 3033-3044.
- (142) Fontaine, F.; Pastor, M.; Sanz, F. Incorporating molecular shape into the alignment-free Grid-Independent Descriptors. *J. Med. Chem.* **2004**, *47*, 2805-2815.
- (143) Stiefl, N.; Bringmann, G.; Rummey, C.; Baumann, K. Evaluation of extended parameter sets for the 3D-QSAR technique MaP: Implications for interpretability and model quality exemplified by antimalarially active naphthylisoquinoline alkaloids. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 347-365.
- (144) Stiefl, N.; Baumann, K. Structure-based validation of the 3D-QSAR technique MaP. *J. Chem. Inf. Model.* **2005**, *45*, 739-749.
- (145) Sanner, M. F.; Olson, A. J.; Spehner, J. C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **1996**, *38*, 305-320.
- (146) Pascual-Ahuir, J. L.; Silla, E. GEPOL: An Improved Description of Molecular Surfaces. I. Building the Spherical Surface Set. *J. Comput. Chem.* **1990**, *11*, 1047-1060.
- (147) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509-10524.
- (148) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual high throughput screen by 4D-QSAR analysis: Application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151-1160.
- (149) Liu, J.; Pan, D.; Tseng, Y.; Hopfinger, A. J. 4D-QSAR analysis of a series of antifungal p450 inhibitors and 3D-pharmacophore comparisons as a function of alignment. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2170-2179.
- (150) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics* **1999**, *9*, 477-489.
- (151) Krasowski, M. D.; Hong, X.; Hopfinger, A. J.; Harrison, N. L. 4D-QSAR analysis of a set of propofol analogues: mapping binding sites for an anesthetic phenol on the GABA(A) receptor. *J. Med. Chem.* **2002**, *45*, 3210-3221.
- (152) Hong, X.; Hopfinger, A. J. 3D-pharmacophores of flavonoid binding at the benzodiazepine GABA(A) receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324-336.
- (153) Santos, O. A.; Hopfinger, A. J. The 4D-QSAR paradigm: Application to a novel set of nonpeptidic HIV protease inhibitors. *Quant. Struct.-Act. Relat.* **2002**, *21*, 369-381.
- (154) Vedani, A.; Briem, H.; Dobler, M.; Dollinger, H.; McMasters, D. R. Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *J. Med. Chem.* **2000**, *43*, 4416-4427.

- (155) Vedani, A.; Dobler, M. 5D-QSAR: the key for simulating induced fit? *J. Med. Chem.* **2002**, *45*, 2139-2149.
- (156) Vedani, A.; Dobler, M.; Lill, M. A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* **2005**, *48*, 3700-3703.
- (157) Whittle, M.; Willett, P.; Klaffke, W.; van Noort, P. Evaluation of similarity measures for searching the dictionary of natural products database. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 449-457.
- (158) 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.
- (159) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435-442.
- (160) 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.
- (161) 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.
- (162) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422-1426.
- (163) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793-806.
- (164) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134-1146.
- (165) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Design* **2000**, *20*, 1-16.
- (166) Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. Prediction of the n-octanol/water partition coefficient, logP, using a combination of semiempirical MO-calculations and a neural network. *J. Mol. Model.* **1997**, *3*, 142-155.
- (167) Fontana, P.; Pretsch, E. Automatic spectra interpretation, structure generation, and ranking. In *J. Chem. Inf. Comput. Sci.*, 2002; pp 614-619.
- (168) Meiler, J.; Maier, W.; Will, M.; Meusinger, R. Using neural networks for C-13 NMR chemical shift prediction- comparison with traditional methods. *J. Magn. Reson.*, 2002; pp 242-252.
- (169) Baumann, K.; Clerc, J. T. Computer-assisted IR spectra prediction - Linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997**, *348*, 327-343.
- (170) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192-204.
- (171) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery - 1. Applications of genetic algorithms to the prediction of blood-brain partitioning of a large set of drugs. *J. Mol. Model.* **2002**, *8*, 337-349.
- (172) Floriano, W. B.; Vaidehi, N.; Zamanakos, G.; Goddard, W. A., 3rd HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J. Med. Chem.* **2004**, *47*, 56-71.
- (173) Kubinyi, H. Hydrogen Bonding: The Last Mystery in Drug Design? *Pharmacokinetic Optimization in Drug Research. Biological, Physicochemical, and Computational Strategies*; Helvetica Chimica Acta and Wiley-VCH: Zurich, 2001; pp 513-524.
- (174) Hill, R. A.; Kirk, D. N.; Makin, H. L. J.; Murphy, G. M. *Dictionary of Steroids*; Chapman&Hill: London.
- (175) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335-373.
- (176) Davis, A. M.; Teague, S. J. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew. Chem.-Int. Edit. Engl.* **1999**, *38*, 737-749.
- (177) Morgan, B. P.; Scholtz, J. M.; Ballinger, M. D.; Zipkin, I. D.; Bartlett, P. A. Differential Binding-Energy - a Detailed Evaluation of the Influence of Hydrogen-Bonding and Hydrophobic Groups on the Inhibition of Thermolysin by Phosphorus-Containing Inhibitors. *J. Am. Chem. Soc.* **1991**, *113*, 297-307.
- (178) Doytchinova, I.; Valkova, I.; Natcheva, R. Adenosine A2A receptor agonists: CoMFA-based selection of the most predictive conformation. *SAR QSAR Environ. Res.* **2002**, *13*, 227-235.

- (179) Harpalani, A. D.; Snyder, S. W.; Subramanyam, B.; Egorin, M. J.; Callery, P. S. Alkylamides as inducers of human leukemia cell differentiation: a quantitative structure-activity relationship study using comparative molecular field analysis. *Cancer Res.* **1993**, *53*, 766-771.
- (180) Timofeir, S.; Kurunczi, L.; Schmidt, W.; Simon, Z. Steric and electrostatic effects in dye-cellulose interactions by the MTD and CoMFA approaches. *SAR QSAR Environ. Res.* **2002**, *13*, 219-226.
- (181) Lanig, H.; Utz, W.; Gmeiner, P. Comparative molecular field analysis of dopamine D4 receptor antagonists including 3-[4-(4-chlorophenyl)piperazin-1-ylmethyl]pyrazolo[1,5-a]pyridine (FAUC 113), 3-[4-(4-chlorophenyl)piperazin-1-ylmethyl]-1H-pyrrolo-[2,3-b]pyridine (L-745,870), and clozapine. *J. Med. Chem.* **2001**, *44*, 1151-1157.
- (182) Ragno, R.; Marshall, G. R.; Di Santo, R.; Costi, R.; Massa, S.; Rompei, R.; Artico, M. Antimycobacterial pyrroles: synthesis, anti-Mycobacterium tuberculosis activity and QSAR studies. *Bioorg. Med. Chem.* **2000**, *8*, 1423-1432.
- (183) Huang, X.; Liu, T.; Gu, J.; Luo, X.; Ji, R.; Cao, Y.; Xue, H.; Wong, J. T.; Wong, B. L.; Pei, G.; Jiang, H.; Chen, K. 3D-QSAR model of flavonoids binding at benzodiazepine site in GABAA receptors. *J. Med. Chem.* **2001**, *44*, 1883-1891.
- (184) Horwitz, J. P.; Massova, I.; Wiese, T. E.; Wozniak, A. J.; Corbett, T. H.; Sebolt-Leopold, J. S.; Capps, D. B.; Leopold, W. R. Comparative molecular field analysis of in vitro growth inhibition of L1210 and HCT-8 cells by some pyrazoloacridines. *J. Med. Chem.* **1993**, *36*, 3511-3516.
- (185) Chau, P. L.; Dean, P. M. Electrostatic Complementarity between Proteins and Ligands .1. Charge Disposition, Dielectric and Interface Effects. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 513-525.
- (186) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43-53.
- (187) Klebe, G.; Abraham, U. On the prediction of binding properties of drug molecules by comparative molecular field analysis. *J. Med. Chem.* **1993**, *36*, 70-80.
- (188) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545-552.
- (189) Kellogg, G. E.; Abraham, D. J. Hydrophobicity: is LogP(o/w) more than the sum of its parts? *Eur. J. Med. Chem.* **2000**, *35*, 651-661.
- (190) Pajeva, I.; Wiese, M. Molecular modeling of phenothiazines and related drugs as multidrug resistance modifiers: a comparative molecular field analysis study. *J. Med. Chem.* **1998**, *41*, 1815-1826.
- (191) Carrupt, P. A.; Testa, B.; Gaillard, P. *Reviews in Computational Chemistry*; Wiley-VCH: New York, 1997; pp 241-315.
- (192) Bohm, H.-J.; Klebe, G. What can we learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs? *Angew. Chem.-Int. Edit. Engl.* **1996**, *35*, 2588-2614.
- (193) Gohlke, H.; Klebe, G. DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.* **2002**, *45*, 4153-4170.
- (194) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38*, 2681-2691.
- (195) Constans, P.; Hirst, J. D. Nonparametric regression applied to quantitative structure-activity relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 452-459.
- (196) Ren, S. Modeling the toxicity of aromatic compounds to tetrahymena pyriformis: the response surface methodology with nonlinear methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1679-1687.
- (197) Zheng, W. F.; Tropsha, A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185-194.
- (198) Goel, M.; Jain, D.; Kaur, K. J.; Kenoth, R.; Maiya, B. G.; Swamy, M. J.; Salunke, D. M. Functional equality in the absence of structural similarity: an added dimension to molecular mimicry. *J. Biol. Chem.* **2001**, *276*, 39277-39281.
- (199) Schlotter, K.; Boeckler, F.; Hubner, H.; Gmeiner, P. Fancy bioisosteres: Metallocene-derived G-protein-coupled receptor ligands with subnanomolar binding affinity and novel selectivity profiles. *J. Med. Chem.* **2005**, *48*, 3696-3699.
- (200) Brown, K. A.; Howell, E. E.; Kraut, J. Long-range structural effects in a second-site revertant of a mutant dihydrofolate reductase. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 11753-11756.

- (201) Lukacova, V.; Balaz, S. Multimode ligand binding in receptor site modeling: implementation in CoMFA. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2093-2105.
- (202) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1-12.
- (203) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579-586.
- (204) Hawkins, D. M.; Basak, S. C.; Shi, X. QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663-670.
- (205) Robertson, S. E.; Jones, K. S. Relevance Weighting of Search Terms. *J. Am. Soc. Inf. Sci.* **1976**, *27*, 129-146.
- (206) Aitchison, J.; Aitken, C. G. G. Multivariate binary discrimination by the kernel method. *Biometrika* **1976**, *63*, 413-420.
- (207) 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.
- (208) Scitegic, Inc., San Diego, CA. - <http://www.scitegic.com>.
- (209) PipelinePilot 5.1, available from Scitegic. <http://www.scitegic.com/>.
- (210) Xia, X. Y.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463-4470.
- (211) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256-3266.
- (212) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708-1718.
- (213) Chen, B. N.; Harrison, R. F.; Hert, J.; Mpanhanga, C.; Willett, P.; Wilton, D. J. Ligand-based virtual screening using binary kernel discrimination. *Mol. Simul.* **2005**, *31*, 597-604.
- (214) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design: An Introduction*; Wiley-VCH: Weinheim, 1999.
- (215) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. 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. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 11322-11326.
- (216) Sternberg, M. J. E.; Muggleton, S. H. Structure activity relationships (SAR) and pharmacophore discovery using Inductive Logic Programming (ILP). *QSAR Comb. Sci.* **2003**, *22*, 527-532.
- (217) von Korff, M.; Steger, M. GPCR-tailored pharmacophore pattern recognition of small molecular ligands. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1137-1147.
- (218) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273-297.
- (219) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: Berlin, 1995.
- (220) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5-14.
- (221) Muller, K. R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'drug-likeness' with Kernel-based learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 249-253.
- (222) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048-2056.
- (223) Trotter, M. W. B.; Holden, S. B. Support vector machines for ADME property classification. *QSAR Comb. Sci.* **2003**, *22*, 533-548.
- (224) 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.
- (225) Weston, J.; Perez-Cruz, F.; Bousquet, O.; Chapelle, O.; Elisseeff, A.; Scholkopf, B. Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics* **2003**, *19*, 764-771.
- (226) Labute, P. Binary QSAR: a new method for the determination of quantitative structure-activity relationships. *Pac. Symp. Biocomput.*, 1999; pp 444-455.
- (227) Faulon, J. L.; Churchwell, C. J.; Visco, D. P., Jr. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721-734.

- (228) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H. W.; Town, W. G.; Yapp, A. M. Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **1973**, *13*, 153 - 157.
- (229) Clark, M.; Cramer, R. D.; Vanopdenbosch, N. Validation of the General-Purpose Tripos 5.2 Force-Field. *J. Comput. Chem.* **1989**, *10*, 982-1012.
- (230) Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81-106.
- (231) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, 1997.
- (232) Domingos, P.; Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* **1997**, *29*, 103-130.
- (233) Harper, G. The Selection of Compounds for Screening in Pharmaceutical Research; University of Oxford, 1999.
- (234) OpenBabel, <http://openbabel.sourceforge.net/>.
- (235) Brown, R. D.; Martin, Y. C. Use of structure Activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572-584.
- (236) Cushman, D. W.; Ondetti, M. A. Design of angiotensin converting enzyme inhibitors. *Nat. Med.* **1999**, *5*, 1110-1113.
- (237) 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.
- (238) Natesh, R.; Schwager, S. L.; Sturrock, E. D.; Acharya, K. R. Crystal structure of the human angiotensin-converting enzyme-lisinopril complex. *Nature* **2003**, *421*, 551-554.
- (239) ISIS, Version 2.1.4, Molecular Design Ltd., San Leandro, USA.
- (240) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502-4520.
- (241) Beroza, P.; Villar, H. O.; Wick, M. M.; Martin, G. R. Chemoproteomics as a basis for post-genomic drug discovery. *Drug Discov. Today* **2002**, *7*, 807-814.
- (242) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273-1280.
- (243) Kohavi, R.; Becker, B.; Sommerfield, D. Improving Simple Bayes. *Proc. Europ. Conf. Mach. Learn.* **1997**.
- (244) Moreau, G.; Broto, P. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359-360.
- (245) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures - perception, auto-correlation descriptor and SAR studies - auto-correlation descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66-70.
- (246) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775.
- (247) MDL Drug Data Report; MDL ISIS/HOST software, MDL Information Systems, Inc.
- (248) Concord, Version 4.0.7, Tripos Inc., St. Louis, Minnesota, USA.
- (249) Pearlman, R. S. CONCORD: Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Design Automation News* **1987**, *2*, 5-7.
- (250) Glen, R. C.; A-Razzak, M. Applications of Rule-Induction in the Derivation of quantitative structure-activity relationships. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 349-383.
- (251) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: Ranking, voting, and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536-1548.
- (252) MOE (Molecular Operating Environment); Chemical Computing Group Inc.: Montreal, Quebec, Canada.
- (253) Yamamoto, Y.; Kamiya, K.; Terao, S. Modeling of human thromboxane A2 receptor and analysis of the receptor-ligand interaction. *J. Med. Chem.* **1993**, *36*, 820-825.
- (254) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499-2510.
- (255) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569-6583.
- (256) Vinter, J. G. Extended electron distributions applied to the molecular mechanisms of intermolecular interactions. *J. Comput.-Aided Mol. Des.* **1994**, *9*, 653-668.

- (257) Klamt, A.; Schuurmann, G. Cosmo - a New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc.-Perkin Trans. 2* **1993**, 799-805.
- (258) Eckert, F.; Klamt, A. *COSMOtherm*; Version C1.2, Release 01.04 ed.; COSMOlogic GmbH & Co. KG: Leverkusen, Germany.
- (259) Klamt, A.; Eckert, F.; Hornig, M. COSMO-RS: A novel view to physiological solvation and partition questions. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 355-365.
- (260) Klamt, A. Conductor-Like Screening Model for Real Solvents - a New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224-2235.
- (261) Murray-Rust, P.; Glusker, J. P. Directional Hydrogen-Bonding to Sp²-Hybridized and Sp³-Hybridized Oxygen-Atoms and Its Relevance to Ligand Macromolecule Interactions. *J. Am. Chem. Soc.* **1984**, *106*, 1018-1025.
- (262) Tomasi, J.; Persico, M. Molecular-Interactions in Solution - an Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.* **1994**, *94*, 2027-2094.
- (263) Schafer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. COSMO Implementation in TURBOMOLE: Extension of an efficient quantum chemical code towards liquid systems. *PCCP Phys. Chem. Chem. Phys.* **2000**, *2*, 2187-2193.
- (264) Ahlrichs, R.; Bar, M.; Haser, M.; Horn, H.; Kolmel, C. Electronic-Structure Calculations on Workstation Computers - the Program System Turbomole. *Chem. Phys. Lett.* **1989**, *162*, 165-169.
- (265) Klamt, A. *COSMO-RS, From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier: Amsterdam, 2005.
- (266) Hornig, M.; Klamt, A. COSMOfrag: A Novel Tool for High Throughput ADME Property Prediction and Similarity Screening Based on Quantum Chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1169 -1177.
- (267) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic 3-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008.
- (268) Elowe, N. H.; Blanchard, J. E.; Cechetto, J. D.; Brown, E. D. Experimental screening of dihydrofolate reductase yields a "test set" of 50,000 small molecules for a computational data-mining and docking competition. *J. Biomol. Screen.* **2005**, *10*, 653-657.
- (269) Zolli-Juran, M.; Cechetto, J. D.; Hartlen, R.; Daigle, D. M.; Brown, E. D. High throughput screening identifies novel inhibitors of Escherichia coli dihydrofolate reductase that are competitive with dihydrofolate. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2493-2496.
- (270) Feher, M.; Deretey, E.; Roy, S. BHB: a simple knowledge-based scoring function to improve the efficiency of database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1316-1327.
- (271) Jain, A. N. Ligand-based structural hypotheses for virtual screening. *J. Med. Chem.* **2004**, *47*, 947-961.
- (272) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195-209.
- (273) Pan, Y. P.; Huang, N.; Cho, S.; MacKerell, A. D. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267-272.
- (274) 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.
- (275) Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469-474.
- (276) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M. K.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem* **2006**, (ASAP Article, DOI 10.1021/jm050362n).
- (277) Marsden, P. M.; Puvanendrapillai, D.; Mitchell, J. B. O.; Glen, R. C. Predicting protein-ligand binding affinities: a low scoring game? *Org. Biomol. Chem.* **2004**, *2*, 3267-3273.