PERSPECTIVE

OBC
www.rsc.org/obc

Molecular similarity: a key technique in molecular informatics†

Andreas Bender and Robert C. Glen*
Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, United Kingdom CB2 1EW. E-mail: rcg28@cam.ac.uk; Tel: +44 (1223) 336 432

Received 28th June 2004, Accepted 9th September 2004
First published as an Advance Article on the web 14th October 2004

Molecular Informatics utilises many ideas and concepts to find relationships between molecules. The concept of similarity, where molecules may be grouped according to their biological effects or physicochemical properties has found extensive use in drug discovery. Some areas of particular interest have been in lead discovery and compound optimisation. For example, in designing libraries of compounds for lead generation, one approach is to design sets of compounds 'similar' to known active compounds in the hope that alternative molecular structures are found that maintain the properties required while enhancing *e.g.* patentability, medicinal chemistry opportunities or even in achieving optimised pharmacokinetic profiles. Thus the practical importance of the concept of molecular similarity has grown dramatically in recent years. The predominant users are pharmaceutical companies, employing similarity methods in a wide range of applications *e.g.* virtual screening, estimation of absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) and prediction of physicochemical properties (solubility, partitioning *etc.*). In this perspective, we discuss the representation of molecular structure (descriptors), methods of comparing structures and how these relate to measured properties. This leads to the concept of molecular similarity, its various definitions and uses and how these have evolved in recent years. Here, we wish to evaluate and in some cases challenge accepted views and uses of molecular similarity. Molecular similarity, as a paradigm, contains many implicit and explicit assumptions in particular with respect to the prediction of the binding and efficacy of molecules at biological receptors. The fundamental observation is that molecular similarity has a context which both defines and limits its use. The key issues of solvation effects, heterogeneity of binding sites and the fundamental problem of the form of similarity measure to use are addressed.

1. Introduction

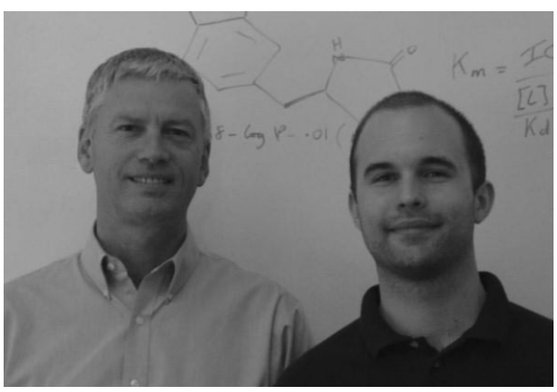
Many "rational" drug design efforts are based on a principle which states that structurally similar compounds are more likely to exhibit similar properties.[1–3] Indeed, the observation that common substructural fragments lead to similar biological activities can be quantified from database analysis.[4,5] 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.[1,6] Similarity methods have found particular favour in the pharmaceutical industry.[7–9] Indeed, medicinal chemistry relies heavily on the concept of bioisosterism in which similar substructures may be interchanged whilst maintaining some degree of activity.[4,10]

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

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

Andreas Bender is a PhD candidate with Robert C. Glen at the Unilever Centre for Molecular Science Informatics, University of Cambridge. He received his BSc in chemistry from the University of Technology (TU), Berlin and, after a year at Trinity College Dublin, he graduated (MSc) from Johann-Wolfgang Goethe University, Frankfurt, where he worked with Gisbert Schneider on the analysis and prediction of mitochondrial transit peptides of P. falciparum. His studies were supported by the German National Merit Foundation (Studienstiftung des Deutschen Volkes) and he is now a member of Darwin College and a Cambridge Gates Scholar. His research interests include the development of new approaches to describe similarity of molecules and their applications such as virtual screening, the prediction of physicochemical properties such as logP, pKa and logD and ADME/Tox prediction.

Robert Glen obtained his PhD in X-Ray Crystallography and organic synthesis from the University of Stirling. One of the highlights was the first co-crystallisation of a reactant and product of a chemical reaction in a single crystal. At the Wellcome Foundation he created the computer-aided molecular design group which included protein crystallography, computational chemistry, molecular transport properties and electrochemistry. He designed the GASP and GOLD computer programs which are used extensively in the pharmaceutical industry, is a co-inventor of Zomig (AstraZeneca) for migraine and invented two other compounds that have entered Phase-2 development. He became Vice President for Collaborative Research at Tripos Inc., assisted in setting up three biotechnology companies, obtained research grants of $3.8M and directed collaborative and contract research in drug discovery. He then moved to Cambridge as Director of the new Unilever Centre for Molecular Science Informatics. He has published over 80 papers and has numerous patents. He is on the SAB of a number of institutes, biotechnology and pharmaceutical companies, is a fellow of the RSC and member of the RSC publications board, an honorary member of the AACR, serves on the Netherlands genomics and bioinformatics initiatives and is an editor of the Encyclopaedia of Computational Chemistry and the advisory board of the Journal of Chemical Informatics and computer science.



Andreas Bender (right) and Robert C. Glen (left)

DOI: 10.1039/b409813g

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 pipeline.[11] 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 2009,[12] 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 methods[13] 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 risk.[14]

To illustrate the steady growth of this area over the last two decades, the number of publications indexed by the Web of Knowledge[15] containing "molecular similarity" in the title or in either the title, abstract or keywords is shown in Fig. 1. 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.

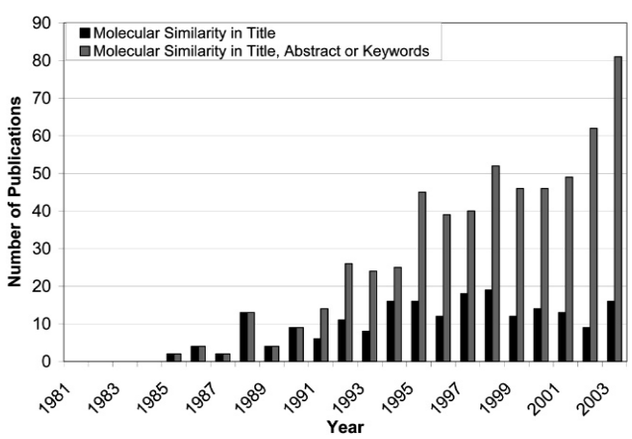


Fig. 1 Number of publications indexed by Web of Knowledge per year which contain "molecular similarity" in the title (black) or in title, abstract or keywords (grey).

Molecular similarity is a dynamic and evolving area of research and has been regularly reviewed. Johnson and Maggiora[1] and Dean[6] wrote comprehensive books in this area. Recently, books by Leach and Gillet[16] and Gasteiger[17] have included sections on molecular similarity. Recent general reviews of molecular similarity are given by Willett *et al.*,[13] Walters *et al.*,[18] Gillet *et al.*[19] and Bajorath,[20] a good critique, particularly of the misuse of similarity measures is given by Nikolova and Jaworska.[21] A justification for the large number of molecular similarity methods is given by Sheridan and Kearsley.[7] Bajorath discusses the role of similarity in the integration of *in silico* and *in vitro* screening,[22] while Johnson *et al.*[23] 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 Kubinyi.[24] Finally the reader is referred to Tversky,[25] 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'.[26] 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 Fig. 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 used.[7] One recent dataset that falls into this category was published by Hert *et al.*.[27] 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 Ki) is often a gross approximation of the ligand binding event being studied; dose response curves are hardly ever collinear with different pharmacodynamics and Ki 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. Also, 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 analysis[28] (CoMFA) or quantum molecular similarity[29,30] 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 alignment.[31–33] Some of these descriptors and analysis methods possess the property of back-projectability of features from descriptor space to geometrical space. Thus, *e.g.* it is possible to generate human-understandable models (*e.g.* "a hydrogen acceptor 7 Å apart from a hydrogen donor is crucial") and also to optimise structures according to the model.

Generating descriptors has a long history[34–36] 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 coefficient[37] (also known as the 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 of this type are known.[13,38,39] Over the last few 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.[40,41] Users of these algorithms should 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 molecules.[42] 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 discusses this in great detail in his well-known paper on neighbourhood behaviour,[2] the essence of which is also illustrated in Fig. 2. Indeed, the question of how similar molecules have to be to display neighborhood behaviour is probably dependant on the property (often biological activity) in question and in particular on the (often assumed) linear behaviour of the relationship between the descriptor and biological activity (and on the variation of the activity itself being a continuous free energy relationship). Martin's study[3,42] demonstrates that compound libraries which commonly contain series of analogues 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 (compounds computed to be similar); obviously they will often find their own analogous series or molecules synthesised specifically to act at that receptor. This skews results for evaluation of virtual screening experiments. This 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 statistics due to unfortunate experimental design. 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 *e.g.* by selecting a set of structures applied to which the algorithm gives best results. Similarity is, after all,[43] a 'fuzzy' concept.

As in other areas of data analysis, machine learning approaches have been applied extensively to the field of molecular similarity. Among others, ID3,[44] linear learning machine,[45] nonlinear mapping,[46,47] support vector machines (SVMs),[48] binary kernel classification[49,50] and the Bayesian classifier.[43,51–53]

We shall give an overview of molecular descriptors used in similarity calculations in chapter 2, followed by a criticism of one of the important aspects of similarity calculations in chapter 3: that similarity possesses a context; it is determined by the environment within which it is perceived (or computed). Chapters 4 and 5 discuss two possible pitfalls with similarity calculations: that desolvation effects are often not additive in a simple fashion and are generally neglected (and electrostatic complementarity is overemphasized) and that linear methods are not capable of modelling a relationship between descriptors and activity if the underlying activity space is rugged. Recent developments in descriptor development are discussed in chapter 6 while chapter 7 deals with implicit properties of binary representations of molecules. Applications of machine learning methods in similarity calculations are given in chapter 8, before we conclude and give an outlook in chapter 9.

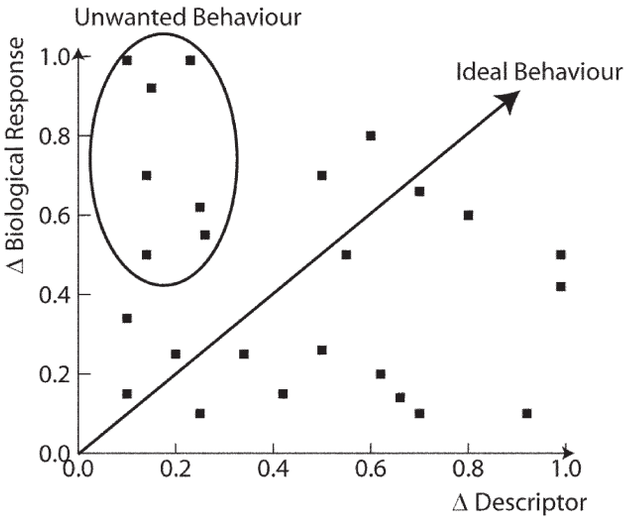


Fig. 2 Illustration of neighbourhood behaviour which shows the relationship between the change in biological activity for pairs of a series of compounds plotted against the change in the descriptor for these pairs of compounds. The ideal behaviour would result in the lower triangle of the plot being occupied, with the most favourable outcome being that small changes in descriptor space are related to small changes in biological response (the typical medicinal chemistry paradigm). Small or large changes in biological response with large changes in descriptor space are also a common feature of such analyses. However, an unsuitable descriptor should not have the property that a large change in biological response results from small changes in the descriptor.

2. 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.

Descriptors can be broken down into those which are derived from the molecular graph, those which depend on molecular shape (conformation) and those which in addition require calculation of the molecular wavefunction. Added to this are descriptors which describe modifications of the molecular structure (*e.g.* ionisation, pK_a) or derivations of surrogate experimental measurements (*e.g.* log *P*).

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 databases[54] and database comparisons (distinguishing between drugs/non-drugs [non-drugs, an interesting discussion in itself][48,55–57]). It should be noted that, in practice, one commonly assigns a large number of one-dimensional descriptors for similarity calculations because the relevant pro-

perty is in most cases either unknown or can only be represented as a combination of different one-dimensional properties.

One-dimensional linear representations attempt to represent the molecule as a linear string or tree 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 molecules.[58] An overview of methods to derive linear molecular descriptors is given by Baumann.[59]

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.[60,61] As yet, there are hardly any applications known using third-generation topological indices.[62] Several hundred alternative topological descriptors have been published.[63–65] 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 Balaban[66] and Estrada and Uriarte.[62] Kier and Hall[67] 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".[68]

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.*,[69] 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 ref. 70. 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 type.[71,72] This is still an active area of research.[19,73] Ghose developed counts of a selected set of 110 fragments based on carbon, hydrogen, oxygen, nitrogen, sulfur and halogens as descriptors[74] which were initially applied to the prediction of log *P*. Rarey and Dixon[75] represent molecules as one-dimensional, potentially branched, sequences designated "feature trees". Other examples for fragment-based descriptors use reduced graphs,[70,76] "molecular tree" fingerprints[77,78] or related "atom environments".[51,52,79,80] "Mini fingerprints" also contain bits which denote the presence or absence of fragments.[81–83] For both "molecular tree" fingerprints and atom environments, fragments spanning two bonds from a central atom (five atoms in diameter) were found to be most effective in similarity searching[51] and QSAR studies.[78]

Descriptors derived from combinations of other descriptors, using correlation or principal components methods are also popular. BCUT descriptors[84,85] 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 (an interesting concept in itself: similarity can at least be quantified, 'diversity' is as big as you want to make it!) in compound library design is desired.

Another class of field-based descriptors are derived from the effects of a test probe at a distance. This could be *e.g.* a positive charge, a water molecule or a methyl probe. Generally, additional properties such as atom centred partial charges have to be pre-computed before calculation of the field effect. This group of descriptors differs fundamentally from the previous group in that they use three-dimensional information of a molecule for their derivation. Because of the number of data points (often termed "grid points" based on a regular grid) that are necessary for a sensible resolution, they are often computationally more demanding than two-dimensional methods. Field-based descriptors generally require alignment of the molecules to be compared. This is trivial only in the case of compound analogues and of course makes the assumption of near identical competitive binding. Many different methods have been developed in this area with the broad separation being between quantum-mechanical methods and non-quantum mechanical methods. Quantum similarity was introduced in the early 1980's.[29] Hodgkin and Richards[86] later introduced a related index that took into account not only electron distribution but also electron density. Walker *et al.*[87] and Good *et al.*[88,89] 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 shape.[90] For a review on quantum similarity, see Carbo-Dorca and Besalu,[30] and latest developments see refs. 91–93. For a basic introduction to the subject see ref. 94 and for a discussion of the significance of QM methods in similarity (particularly atoms in molecules theory) see Nikolova and Jaworska[21] and Boon *et al.*.[93] 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 wavefunction. 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.

Methods based on grid based descriptors, many of which owe their inspiration to Goodford's work in field based methods[95] in combination with a robust statistical method for variable selection (partial least squares, PLS),[96] were introduced in the late 1980's with the comparative molecular field method, CoMFA.[28] This method was also the basis of Klebe *et al.*'s comparative molecular similarity analysis (CoMSIA) approach.[97,98]

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 hypothesis[99] and in many applications it involves identifying key (similar) 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, although it would often be assumed that alignment of key interaction points would occur in the ligand/receptor interaction). They are often referred to as multiple-point-pharmacophores: two-point pharmacophores[100,101] (2PP), which are known as atom pairs and represent all possible pairs of atoms in the molecule, three-point pharmacophores[102–107] (3PP), which allow for a more detailed representation of interatomic distances, and four-point pharmacophores[108,109] (4PP), which are able to distinguish between geometric isomers. Four-point pharmacophores are discussed later.

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 properties.[110] Some examples of the utility of surface based descriptors follow. Gaillard *et al.*[111] devised a method to describe molecular lipophilicity potential and validated it by predicting log *P* values. Stanton and Jurs[112] introduced the concept of "charged partial surface area structural descriptors". Jain *et al.*'s Compass method[113] 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 Icepick,[114] 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"[115] 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 hashkeys of the molecular surface, compared to a panel of reference compounds.[116] 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 retrieval[117] 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.*[118] and shortly afterwards followed by in their *in silico* counterparts.[119,120] The latter were, for example, employed in library design.[121] For a recent review see Briem and Lessel.[122]

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 Wilkins,[123] who used molecular transforms to calculate the diffraction pattern from an X-ray derived three-dimensional structure. Electron diffraction was also used in the 3D-MoRSE (molecule representation of structures based on electron diffraction) approach.[124] The first descriptor calculated from the vibrational spectra of molecules is the EVA descriptor.[125] Here, fundamental frequencies of the vibrational spectrum are calculated and used for the comparison of molecules. A different approach[126] defines fuzzy peak areas to derive molecular features from an infrared spectrum, followed by principal components analysis. Although spectra are a "natural" way to convert a molecule into a one-dimensional representation, small changes often introduce major changes in the spectrum (*e.g.* bond lengths) and the representation in descriptor space. These changes often make it difficult to use this approach consistently as a similarity index.

3. 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[79,127] or NMR[128,129] or IR[130] spectra. However, of particular importance for the pharmaceutical industry are properties like absorption, distribution, metabolism, excretion and toxicity (ADME/Tox)[131,132] and bioactivity.[8,51,122,133] These depend to a great extent on perturbations introduced by specific interactions between ligand and receptor (or transporter molecule, as in the case of active transport in absorption, *etc.*). 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–amine 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 replacement.[134] 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.*.[134,135] 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 by *e.g.* fingerprint based Tanimoto similarity indices, 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 properties.[135] There are no absolute measures of molecular similarity and each case requires the selection of appropriate properties and classification methods.

4. Omitting important features: Free energy, enthalpy and entropic descriptions for binding need 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 Fig. 3. The pharmacodynamics (on–off) rates of the ligand are also of importance therefore the stability of the complex strongly influences the bioactivity observed.

It is commonly accepted that the energy gained during ligand–receptor complex formation by removing water from lipophilic surfaces is largely responsible for overall stability of the complex formed.[136] 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 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. 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.[137,138] 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)[28] 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 1.

In order to examine predictivity using electrostatic information, Chau and Dean studied electrostatic complementarity of 34 ligand–receptor complexes.[139] Calculating correlation coefficients of van der Waals surface points, they found significant correlation 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 dependent 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 recognition.[140]

Klebe and Abraham[141] discuss the influence of enthalpic and entropic factors on binding. Building CoMFA models for renin inhibitors, they conclude 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 ($\Delta G° = \Delta H° - T\Delta S°$), 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"[141] 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 log *P* which in these instances may have more predictive value. This is understandable, since log *P* 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 (see Fig. 3). Soon after publication of CoMFA,[28] a third, hydrophobic field was introduced.[142] 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

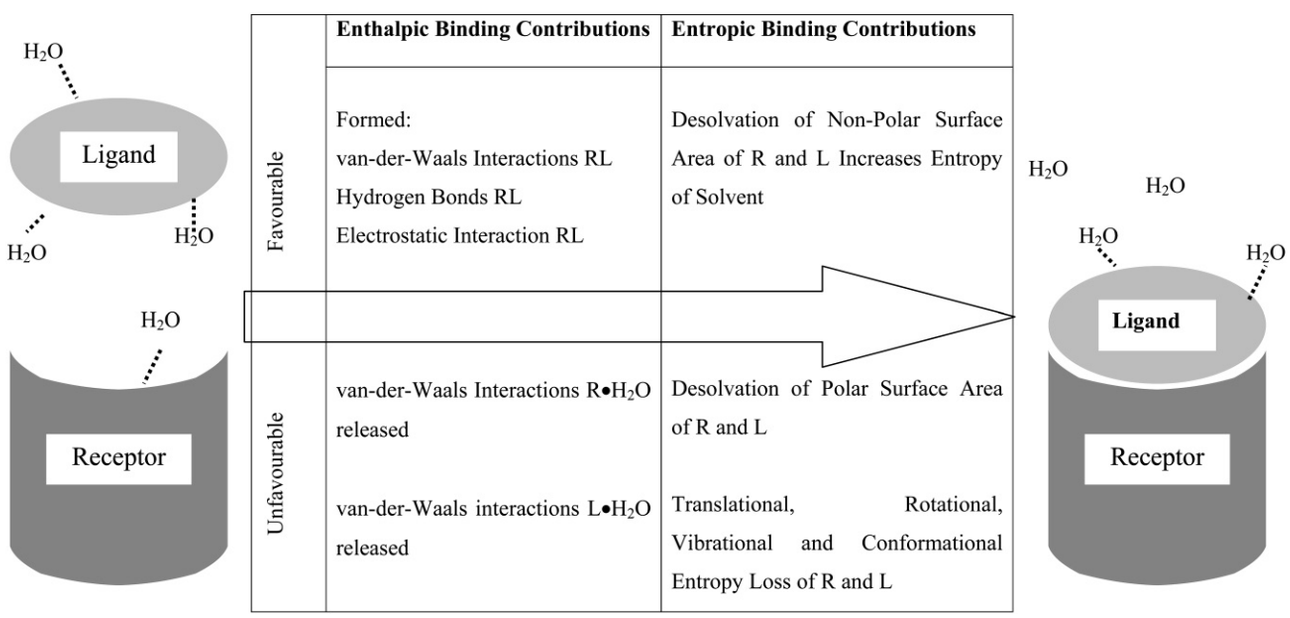


Fig. 3 Enthalpic and entropic contributions to ligand (L)–receptor (R) binding. It is commonly assumed that non-polar surface areas contribute to stability of the ligand–receptor complex through desolvation energies while polar surface areas are important for selectivity.

Table 1 Importance of steric and electrostatic field components in applications of comparative molecular field analysis (CoMFA) [Cramer 1988].[28] Interestingly, electrostatic and steric effects are often seen to contribute (or not) in significantly varying degrees to CoMFA models (see chapter 4 for discussion)

Steric > Electrostatic	Steric ~ Electrostatic	Electrostatic > Steric
Selection of the most predictive conformation of adenosine A2A receptor agonists[209]	Alkylamides as inducers of human leukemia cell differentiation[211]	Affinity of dyes for cellulose fiber[215]
Dopamine D4 receptor antagonists[210]	SAR of antifungal pyrrole derivatives[212]	
	Flavonoids binding at benzodiazepine site in GABAA receptors[213]	
	Cytotoxicity of substituted acridines against HCT-8 cell line relative to mouse leukaemia L1210 cells[214]	

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 non progressive interactions 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 interpret[143] 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 reversal.[144]

The recently published HINT (hydrophobic interaction) force field attempts to include hydrophobic information, derived from log *P* data.[143] Other molecular descriptors, such as GRIND[31] or CoMSIA[97,98] also make use of hydrophobic fields, as does molecular lipophilic potential (MLP).[145]

To summarize, most commonly employed molecular similarity methods, such as in the original CoMFA,[28] in which ligands are compared to each other, often appear to work well without taking any of the effects listed in Fig. 3 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 Klebe,[146] Gohlke and Klebe[147] and Brooijmans and Kuntz.[136]

5. Local similarity requires non-linear models. Or: Why linear regression techniques are not the method of choice

It is the first step of molecular similarity methods to represent molecules in "chemical space". Feature selection is the (optional) second step; comparison of structures, either one-on-one or by comparing each library structure to some kind of consensus model from multiple query structures, is the mandatory last step. Information from multiple molecules may be combined in the model generation step in order to improve predictive capabilities. As a necessary condition, statistical properties of the derived model have to be satisfactory; as an advantage, the prediction rules derived may (ideally) be understandable by humans.

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 method[148] 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 superior.[149] 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 Ren.[150] 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. Additional progress in tackling this problem can be seen in the Compass method[113] which also employs a new type of neural network using surface points as descriptors.

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 known.[151] 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 the binding mode of the DHFR inhibitor methotrexate, compared to the binding mode of the natural substrate.[152] In this case, the binding mode was completely inverted following slight changes of molecular structure. To address one aspect of this, multimode ligand binding was implemented into CoMFA[153] and appears to give better results than standard CoMFA.

Using 3D information, popular methods, such as CoMFA,[28] 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 difficult.[154] Alignment of very dissimilar structures (different scaffolds) is not a trivial task. This is illustrated in Fig. 4 and Fig. 5. While alignment of two angiotensin converting enzyme inhibitors of very similar size is possible without problems (Fig. 4), alignment of two structures of different size, which still both show the inhibition of ACE *in vivo*, gives close to arbitrary alignment (Fig. 5).

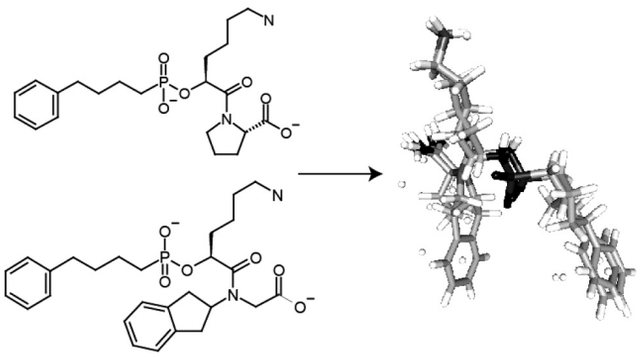


Fig. 4 Alignment of two inhibitors of angiotensin converting enzyme. Both molecules are of comparable size so alignment is, although still time consuming, feasible in an unambiguous manner.

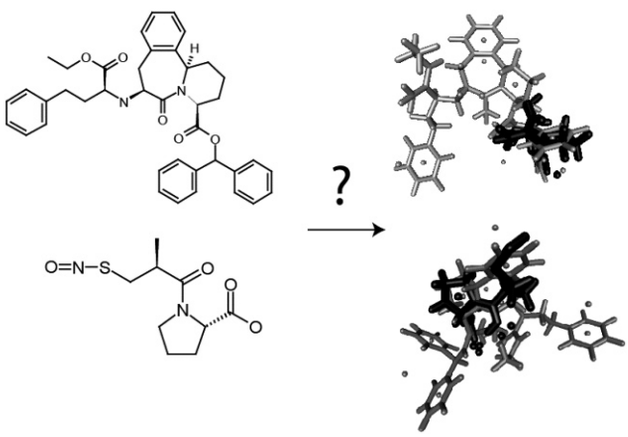


Fig. 5 Alignment of two inhibitors of angiotensin converting enzyme of very different size. Alignment is in this case not feasible in an unambiguous manner. These results were obtained using a genetic algorithm (GASP) for alignment but this problem exists independent of the particular algorithm that is applied.

One novel and interesting way to remove the structural overlay criteria 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 Andrews and Cramer[33,155] 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.

While it may be sensible to apply linear methods to the prediction of physicochemical properties,[112] 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 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 Fig. 6. 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 such as 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.[156–158]

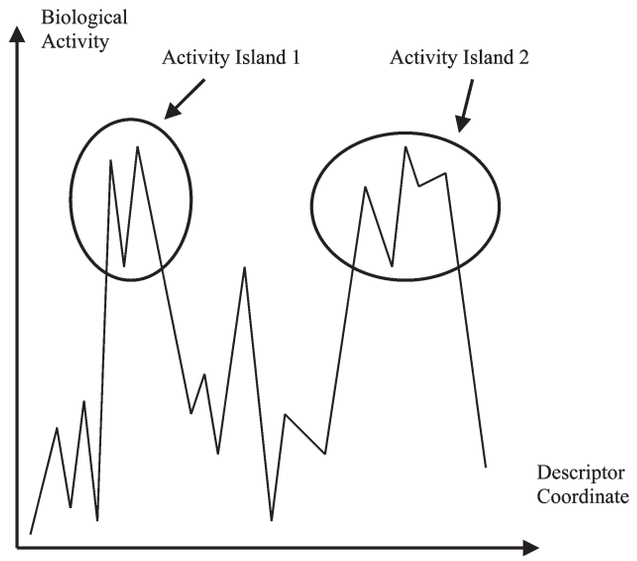


Fig. 6 Variations of a molecular property (here biological activity), depending on the positioning of the structures in property space. Linear regression, which may be appropriate in cases of physicochemical properties such as log *P*, is certainly not appropriate to model this relationship between structure and activity. Methods based on clustering around activity neighbourhoods should be more robust.

6. Recent developments in molecular descriptor generation

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 Consonni.[36]

A descriptor places two molecules in chemical space at a distance that describes their similarity (in this particular descriptor space). The ideal descriptor places two molecules at a distance that is proportional to their difference in bioactivity, physicochemical properties or any other property of the molecules. 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).

New molecular descriptors

As described in the introduction, 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 can be projected back onto the molecules from which it was derived in some chemically meaningful way. However, the solvation effects, as discussed previously, apply in much the same way. These descriptors, 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 original proof of principle steroid 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 Fig. 7 and Fig. 8. In Fig. 7 CoMFA[28] was applied to 21 steroids from the

well-known steroid dataset.[159] The steroids were aligned and using PLS areas in space which contribute favourably or unfavourably 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 Fig. 8 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 classifier.[53] They correspond to experimentally determined binding features[160] and correctly identify the CoA binding pocket and the lipophilic moiety. This information about binding can be used to design novel active entities.

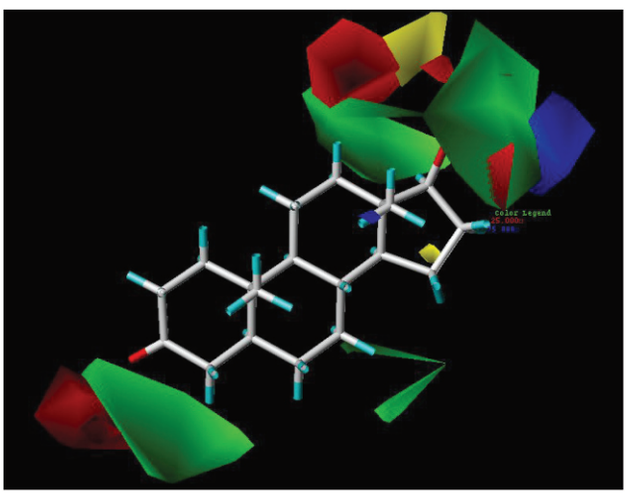


Fig. 7 Areas in space projected back on dihydrotestosterone which are favourable or unfavourable for binding with respect to steric effects (green/yellow) or where negative (blue) or positive (red) charges increase activity.

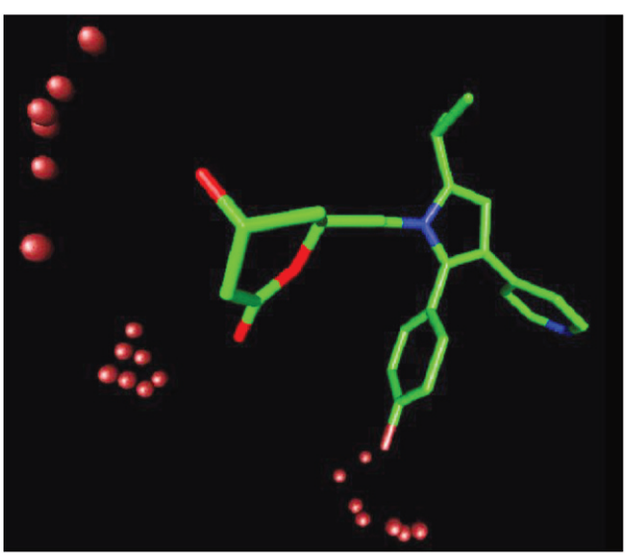


Fig. 8 Features identified as being responsible for binding of a statin to HMG-CoA reductase. Employed were interaction energies of surface points which were encoded in a binary format, information-gain based feature selection and the naïve Bayesian classifier (MOLPRINT method). Features binding to the CoA site as well as to the flexible lipophilic pocket are correctly identified.

The first of the recent descriptors possessing both translational and rotational invariance and back-projectability was published by Pastor *et al.*[31] 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 the maximum product does not correspond to a meaningful feature in every case. At the same time, back-projectability is facilitated. Using a related program, ALMOND,[161] 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 identified.[162] The findings were in agreement with experimental data. More recently, a shape description was implemented as an additional "probe" of the GRID program.[163] 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 statistics.[163]

A related approach was presented by Stiefl *et al.*[32,164] 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 MSMS[165] and others have been found not to give equally spaced surface points so an implementation of the GEPOL algorithm[166] 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 compounds.[32] Again the advantage of translational and rotational invariance as well as back-projectability is seen.

The shape signatures method[167] uses a ray-tracing approach to project a molecule in descriptor space. A ray is reflected inside of a molecule, or alternatively in the binding pocket of a receptor. Then, properties are assigned to each point where the ray is reflected by the surface. Histograms of the distance between reflections as the first dimension and the property at the point of reflection as the second dimension of the descriptor are kept. These histograms can be compared using one of several methods. The histograms, counts of ray lengths between reflections, converge quickly with the number of reflections performed. In the order of 10 000 reflections is still needed, giving rise to computational costs of several minutes per molecule (using P-III/450 machines). The authors present no real application, but preliminary studies using small systems (substituted bicyclic compounds) show intuitively sensible behaviour of the descriptor.

A descriptor similar to the start-end-vector method[168] encodes the topological distance matrix using atom counts,[169] but only encodes the shortest path between atoms that leads to a smaller computational overhead. A geometric extension of the distance matrix also takes three-dimensional information into account and is able to distinguish between different molecular conformations. Overall performance is compared to EVA[125] and

CoMFA[28] and found to be slightly superior in test cases. This may be due to the fact that it considers essentially 2D-information of the molecule which has often been found to be superior to completely 3-dimensional methods (probably due to the additional irrelevant data (noise) introduced when considering 3-D models of molecules).[170] Translational and rotational invariance is found due to the fact that the connectivity table and only relative distance information is used for descriptor generation. Back-projectability is illustrated using examples from the steroid data set but generally found to be cumbersome.

Two complementary approaches are the SIFt (structural interaction fingerprint)[171] and a method published by Putta *et al*..[172] SI fingerprints are a novel method to describe interaction between ligand and receptor that is straightforward and in effect describes 3D interaction information in a one-dimensional bit string. First, all residues which interact with any of the ligands of a ligand library are identified. Then bitstrings are created representing the interaction points of every ligand in the receptor pocket. The Tanimoto coefficient can then be used to compare interaction profiles. If compared to binding modes calculated using PMF and Cscore (a consensus scoring function[173]), it is shown that in particular PMF and less so Cscore, are not able to distinguish between true binding modes and others. The SIFt approach can also be applied for virtual screening by docking ligands into the receptor and calculating interaction fingerprints. These fingerprints can be compared to high affinity ligands at the screening stage. This strategy outperforms Cscore and PMF in this application

The Putta *et al*. method[172] has characteristics in common with SI fingerprints in that it attempts to identify features responsible for binding. Since it only takes into account information from ligand molecules, no receptor information needs to be present. It needs alignment of molecules and sometimes even alignment of key features. Often the overall shape of the ligand is not important, but only subshapes of part of the molecule. One approach to handle this problem is to calculate a small number of terminal points in the query molecule and a larger number of skeleton points in the target molecule.[174] Then subshape matching is possible using triangles inside the molecule. Promising preliminary studies are given orienting NAPAP and a benzamidine fragment, a thorough validation of the method is not yet available.

An explicit method for molecular alignment is presented by Kotani and Higashiura.[175] 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 Meyer and Richards'[176] and Good and Richards'[89] approach are obtained with respect to molecular similarity, but the new method is several orders of magnitude faster.

Extensions of established algorithms: multiple point pharmacophores and multi-dimensional QSAR

Molecular similarity algorithms based on three-point pharmacophores are among the standard algorithms applied in industry.[106] They are based on the three-dimensional structure of the molecule. Interaction types are assigned to all atoms of the molecule before all possible triangles of interaction points are constructed. Several thousand combinations of interaction points and binned triangle distances are possible (*e.g.* about 33000 in the case of six interaction types and ten distance ranges), but only a small fraction of bits is usually occupied (a single digit percentage).

Three-point pharmacophores can also be easily extended to incorporate a fourth pharmacophore point for descriptor generation, giving four-point pharmacophores.[108,177,178] These are able to distinguish between geometrical isomers. A potential shortcoming 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 33000 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 proteases.[177] 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 design,[178] here describing an example combinatorial library based on the Ugi condensation and a serine protease active site.

Nonetheless, the method has not yet found as wide an acceptance in the scientific community as its authors probably intended. This is 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 found,[177] 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.

CLIP (candidate ligand identification program)[179] describes structures by the geometric arrangement of pharmacophore points. A maximum common substructure (MCS) routine based on the Bron–Kerbosch algorithm is employed for matching. This is usually applied to graphs derived from the two-dimensional structure. CLIP finds the largest set of points that is geometrically equivalent in two molecules. In can be used for comparison of molecules and uses fuzzy matching, in sample searches it is reported to perform as well as Unity fingerprints.

Multi-dimensional QSAR approaches are an extension of QSAR algorithms using the three spatial dimensions for descriptor encoding, *e.g.* comparative molecular field analysis (CoMFA). The fourth dimension describes conformational sampling of the molecule.

Originally, four-dimensional QSAR was introduced by Hopfinger *et al*..[180] 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 well as 3D approaches but gives additional information. First, active conformations can be guessed. Second, 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 exist.[109] 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 not necessary, as discussed above. 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 to 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 given.[181] 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 low.[182] Cytochome P450 2D6 inhibitors have already been subject to an earlier 3D/4D QSAR study.[183] A study using Propofol analogues gave identical results concerning the interaction sites.[184] Not surprisingly, analogue compounds are predicted better than structures dissimilar to the training compounds. Flavonoids binding to the BzR site of of GABA-A were examined by Hong and Hopfinger.[185]

Conceptually, the common receptor-independent (RI) 4D QSAR analysis has recently been extended to a receptor-dependent (RD) version.[182] 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 inhibitors.[186]

A modified 4D QSAR algorithm, compared to the Hopfinger *et al*. version, is able to take local induced fit and hydrogen bond flip-flop into account.[187] Allowing for a multiple representation of the receptor another dimension is added, giving a 5D QSAR method.[188] This provides the possibility of allowing induced fit of the ligand–receptor complex.

7. Properties of fingerprints and similarity coefficients, 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.

First, binary representations may not be unbiased representations of molecules. They may 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.

Second, the selection of the similarity coefficient used (of which a large number exist[13]) 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.*[189] (in the context of natural product databases), Salim *et al.*[190] (also covering data fusion) and Holliday *et al.*[39] (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, *e.g.* hydrogen bond donor, allows 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.*[191] has 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.

Size dependence of similarity coefficients

The size dependence of the commonly employed Tanimoto coefficient[37] (Tc) has already been known for several years, favouring larger molecules in similarity and smaller molecules in diversity selections. 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.

Flower[40] 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 more spread out. Thus it should be kept in mind that Tc values are not directly comparable among different queries. The size dependence was also examined by Dixon and Koehler,[41] 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.*[192] 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. Again it is found that the Tc distribution depends on relative bit densities of compounds. Fourteen 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.

Attempting to rid the Tanimoto coefficient of its size bias, a size-modified Tanimoto coefficient was recently introduced.[193] This analysis suggests that using this methodology, size-dependence should be removed. However there exists no comprehen-

sive study of its performance. Interestingly (or confusingly), in a different study[189] its performance was found to be similar to the original Tanimoto coefficient.

Clustering of similarity coefficients and consensus scoring

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 scoring.[39] 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 extended[190] 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),[189] where the Tanimoto coefficient was often outperformed by a factor of two.

The theoretical basis of improving classification results by consensus scoring was presented by Wang and Wang.[194] 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.*,[195] 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 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 functions.[195]

This finding was confirmed when the clustering of similarity coefficients was used to select sets of similarity coefficients for consensus scoring.[39,190] Out of seven targets in all except one (β-lactamase) improvement in classification was observed in most cases when three or more similarity coefficients were used.

There are different possibilities to combine information from several scoring algorithms to provide a single prediction. Using several data sets, Ginn *et al.*[196] 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.

In docking procedures, it has been reported that data fusion improves classification results. This may be due to different scoring functions describing different aspects of the interaction, *e.g.* Poisson–Boltzmann calculations describe the desolvation energy accurately, but do not account for hydrophobic interactions. Reducing the number of false positives may be the principal effect as was seen in analysis of MAP kinase, inosine monophosphate dehydrogenase and HIV protease and 13 scoring functions where data fusion was seen to consistently reduce the number of false positives.[197]

8. 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)[198] 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)[199] 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 approaches,[200] although direct comparison is difficult.

Neural networks have many applications in the literature in the field of clustering similar compounds *e.g.* recently NNs have been applied to profiling of GPCR-active compounds.[201] The hidden aspect of the model produced and the tendency to over fit are the main criticisms of these methodologies. They are particularly good though at 'memorising' a given state, which can be useful when searching for other similar states (similar molecules).

In recent years, support vector machines (SVMs)[202,203] 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.[202,203] 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 inhibitors.[204] In addition, training was fastest of the methods tried. Applied to a drug/non-drug classification problem, SVMs also outperform ANNs slightly.[48] This result is consistent and does not depend on descriptors or size of the datasets. In a similar application predicting drug-likeness and agrochemical-likeness,[205] SVMs consistently outperform ANNs. In addition, a QSAR model that was developed to predict activity outperformed earlier QSAR methods. An application of SVMs for prediction of ADME properties[206] 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 learning.[207] 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 classifier.[208] 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[49,50] and the naïve Bayesian classifier.[51–53] 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 two-dimensional and three-dimensional structure.[51,52]

9. Conclusions and outlook

In this perspective we have reviewed progress in molecular similarity methods and applications and highlighted some of the more challenging problems and assumptions.

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.

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.

Back-projectable descriptors (compared to descriptors without this property) 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.

As understanding of the chemistry and biology of drug action improves and a greater ability to model the underlying mechanisms appears, the need for ‘similarity’ approaches will diminish.

Acknowledgements

We would like to thank Unilever and the Gates Cambridge Trust for funding. A number of anonymous referees are thanked for providing valuable input on the manuscript.

References

1 M. A. Johnson and G. M. Maggiora, in *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York, 1990.
2 D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger, *J. Med. Chem.*, 1996, **39**, 3049.
3 Y. C. Martin, J. L. Kofron and L. M. Traphagen, *J. Med. Chem.*, 2002, **45**, 4350.
4 R. P. Sheridan, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 103.
5 R. P. Sheridan, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1037.
6 P. M. Dean, in *Molecular Similarity in Drug Design*, Kluwer Academic Publishers, Dordrecht, 1994.
7 R. P. Sheridan and S. K. Kearsley, *Drug Discovery Today*, 2002, **7**, 903.
8 H.-J. Bohm and G. Schneider, *Virtual Screening for Bioactive Molecules*, Wiley-VCH, Weinheim, 2000.
9 G. Schneider and H. J. Bohm, *Drug Discovery Today*, 2002, **7**, 64.
10 A. Schuffenhauer, V. J. Gillet and P. Willett, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 295.
11 J. A. DiMasi, R. W. Hansen and H. G. Grabowski, *J. Health Econ.*, 2003, **22**, 151.
12 http://wwwdb.europarl.eu.int/oeil/oeil_ViewDNL.ProcedureView?lang=2&procid=4179’.
13 P. Willett, J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983.
14 N. Greene, *Adv. Drug Deliv. Rev.*, 2002, **54**, 417.
15 Thomson ISI Web of Knowledge—http://www.isinet.com.
16 A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Kluwer Academic Publishers, Dordrecht, 2003.
17 J. Gasteiger, *Handbook of Chemoinformatics*, Wiley-VCH, Weinheim, 2003.
18 W. P. Walters, M. T. Stahl and M. A. Murcko, *Drug Discovery Today*, 1998, **3**, 160.
19 V. J. Gillet, D. J. Wild, P. Willett and J. Bradshaw, *Comput. J.*, 1998, **41**, 547.
20 J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 233.
21 N. Nikolova and J. Jaworska, *QSAR Comb. Sci.*, 2004, **22**, 1006.
22 J. Bajorath, *Nat. Rev. Drug Discov.*, 2002, **1**, 882.
23 M. Johnson, S. Basak and G. Maggiora, *Math. Comput. Modell.*, 1988, **11**, 630.
24 H. Kubinyi, *J. Braz. Chem. Soc.*, 2002, **13**, 717.
25 A. Tversky, *Psychol. Rev.*, 1977, **84**, 327.
26 R. L. Goldstone, in *MIT encyclopaedia of the cognitive sciences*, ed. R. A. Wilson and F. C. Keil, MIT Press, Cambridge, MA, 2004.
27 J. Hert, P. Willett and D. J. Wilton, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1177.
28 R. D. Cramer, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.*, 1988, **110**, 5959.
29 R. Carbo, L. Leyda and M. Arnau, *Int. J. Quantum Chem.*, 1980, **17**, 1185.
30 R. Carbo-Dorca and E. Besalu, *J. Mol. Struct., THEOCHEM*, 1998, **451**, 11.
31 M. Pastor, G. Cruciani, I. McLay, S. Pickett and S. Clementi, *J. Med. Chem.*, 2000, **43**, 3233.
32 N. Stiefl and K. Baumann, *J. Med. Chem.*, 2003, **46**, 1390.
33 R. D. Cramer, *J. Med. Chem.*, 2003, **46**, 374.
34 W. E. Brugger, A. J. Stuper and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 1976, **16**, 105.
35 R. C. Glen and V. S. Rose, *J. Mol. Graphics*, 1987, **5**, 79.
36 R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
37 P. Jaccard, *Bull. Soc. Vaudoise Sci. Nat.*, 1901, **37**, 547.
38 G. W. Adamson and J. A. Bush, *J. Chem. Inf. Comput. Sci.*, 1975, **15**, 55.
39 J. D. Holliday, C. Y. Hu and P. Willett, *Comb. Chem. High-Throughput Screening*, 2002, **5**, 155.
40 D. R. Flower, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 379.
41 S. L. Dixon and R. T. Koehler, *J. Med. Chem.*, 1999, **42**, 2887.
42 Y. C. Martin, in *Euro QSAR 2002: Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, ed. M. G. Ford, D. J. Livingstone, J. Dearden and H. van de Waterbeemd, Prous Science Publishers, Barcelona, 2003.
43 J. W. Godden and J. Bajorath, *QSAR Comb. Sci.*, 2003, **22**, 487.
44 R. C. Glen and M. A. Razzak, *J. Comput.-Aided Mol. Des.*, 1992, **6**, 349.
45 K. S. Jandu, V. Barrett, M. Brockwell, D. Cambridge, D. R. Farrant, C. Foster, H. Giles, R. C. Glen, A. P. Hill, H. Hobbs, A. Honey, G. R. Martin, J. Salmon, D. Smith, P. Woollard and D. L. Selwood, *J. Med. Chem.*, 2001, **44**, 681.
46 D. J. Livingstone, G. Hesketh and D. Clayworth, *J. Mol. Graphics*, 1991, **9**, 115.
47 D. K. Agrafiotis and H. Xu, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 15869.
48 E. Byvatov, U. Fechner, J. Sadowski and G. Schneider, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1882.
49 J. Aitchison and C. G. G. Aitken, *Biometrika*, 1976, **63**, 413.
50 G. Harper, J. Bradshaw, J. C. Gittins, D. V. S. Green and A. R. Leach, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1295.
51 A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 170.
52 A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1708.
53 A. Bender, H. Y. Mussa, G. S. Gill and R. C. Glen, *J. Med. Chem.*, 2004, DOI: 10.1021/jm049611i.
54 G. M. Downs, P. Willett and W. Fisanick, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1094.
55 J. Sadowski and H. Kubinyi, *J. Med. Chem.*, 1998, **41**, 3325.
56 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, 1997, **23**, 3.

57 C. A. Lipinski, *J. Pharmacol. Toxicol. Methods*, 2000, **44**, 235.
58 S. L. Dixon and K. M. Merz, Jr., *J. Med. Chem.*, 2001, **44**, 3795.
59 K. Baumann, *TrAC, Trends Anal. Chem.*, 1999, **18**, 36.
60 A. T. Balaban and T. S. Balaban, *J. Chim. Phys. Phys.-Chim. Biol.*, 1992, **89**, 1735.
61 A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 398.
62 E. Estrada and E. Uriarte, *Curr. Med. Chem.*, 2001, **8**, 1573.
63 C. L. Wilkins and M. Randic, *Theor. Chim. Acta*, 1980, **58**, 45.
64 M. Randic and C. L. Wilkins, *J. Chem. Inf. Comput. Sci.*, 1979, **19**, 31.
65 A. T. Balaban, *Chem. Phys.*, 1982, **89**, 399.
66 A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 339.
67 L. H. Hall and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 1039.
68 G. E. Kellogg, L. B. Kier, P. Gaillard and L. H. Hall, *J. Comput.-Aided Mol. Des.*, 1996, **10**, 513.
69 M. M. Cone, R. Venkataraghavan and F. W. McLafferty, *J. Am. Chem. Soc.*, 1977, **99**, 7668.
70 J. M. Barnard, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 532.
71 S. M. Free, Jr. and J. W. Wilson, *J. Med. Chem.*, 1964, **53**, 395.
72 R. D. Cramer, 3rd, G. Redl and C. E. Berkoff, *J. Med. Chem.*, 1974, **17**, 533.
73 G. Rucker and C. Rucker, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 314.
74 A. K. Ghose and G. M. Crippen, *J. Comput. Chem.*, 1986, **7**, 565.
75 M. Rarey and J. S. Dixon, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 471.
76 Y. Takahashi, M. Sukekawa and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 639.
77 J. L. Faulon, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1204.
78 J. L. Faulon, D. P. Visco, Jr. and R. S. Pophale, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 707.
79 L. Xing and R. C. Glen, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 796.
80 L. Xing, R. C. Glen and R. D. Clark, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 870.
81 L. Xue, F. L. Stahura, J. W. Godden and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 394.
82 L. Xue and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 801.
83 L. Xue, J. W. Godden and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 881.
84 F. R. Burden and D. A. Winkler, *J. Med. Chem.*, 1999, **42**, 3183.
85 R. S. Pearlman and K. M. Smith, *Perspect. Drug Discovery Design*, 1998, **9–11**, 339.
86 E. E. Hodgkin and W. G. Richards, *Int. J. Quantum Chem.*, 1987, 105.
87 P. D. Walker, G. A. Arteca and P. G. Mezey, *J. Comput. Chem.*, 1991, **12**, 220.
88 A. C. Good, E. E. Hodgkin and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 188.
89 A. C. Good and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 112.
90 J. A. Grant and B. T. Pickup, *J. Phys. Chem.*, 1995, **99**, 3503.
91 P. Bultinck, T. Kuppens, X. Girones and R. Carbo-Dorca, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1143.
92 E. Besalu, X. Girones, L. Amat and R. Carbo-Dorca, *Acc. Chem. Res.*, 2002, **35**, 289.
93 G. Boon, W. Langenaeker, F. De Proft, H. De Winter, J. P. Tollenaere and P. Geerlings, *J. Phys. Chem. A*, 2001, **105**, 8805.
94 R. Carbo and B. Calabuig, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 600.
95 P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849.
96 H. Wold, in *Encyclopedia of Statistical Sciences (Vol. 6)*, ed. S. Kotz and N. L. Johnson, Wiley, New York, 1985.
97 G. Klebe, U. Abraham and T. Mietzner, *J. Med. Chem.*, 1994, **37**, 4130.
98 G. Klebe, *Perspect. Drug Discovery Design*, 1998, **12**, 87.
99 C. R. Marshall, C. D. Barry, H. E. Bosshard, R. A. Dammkoehler and D. A. Dunn, in The Conformational Parameter in Drug Design: The Active Analog Approach, *Computer-Assisted Drug Design, ACS Symposium Series 112*, ed. E. Olson and A. E. Christoffersen, ACS, Washington DC, USA, 1979.
100 G. Schneider, W. Neidhart, T. Giller and G. Schmid, *Angew. Chem., Int. Edn. Engl.*, 1999, **38**, 2894.
101 R. P. Sheridan, M. D. Miller, D. J. Underwood and S. K. Kearsley, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 128.
102 P. Gund, *Prog. Mol. Subcell. Biol.*, 1977, **5**, 117.
103 G. W. Bemis and I. D. Kuntz, *J. Comput.-Aided Mol. Des.*, 1992, **6**, 607.
104 R. Nilakantan, N. Bauman and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 79.
105 S. D. Pickett, J. S. Mason and I. M. McLay, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 1214.
106 J. S. Mason, A. C. Good and E. J. Martin, *Curr. Pharm. Des.*, 2001, **7**, 567.
107 Y. C. Martin, *J. Med. Chem.*, 1992, **35**, 2145.
108 J. S. Mason, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme and R. F. Labaudiniere, *J. Med. Chem.*, 1999, **42**, 3251.
109 J. S. Duca and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1367.
110 T. Clark, *J. Mol. Graphics*, 2004, **22**, 519.
111 P. Gaillard, P. A. Carrupt, B. Testa and A. Boudon, *J. Comput.-Aided Mol. Des.*, 1994, **8**, 83.
112 D. T. Stanton and P. C. Jurs, *Anal. Chem.*, 1990, **62**, 2323.
113 A. N. Jain, K. Koile and D. Chapman, *J. Med. Chem.*, 1994, **37**, 2315.
114 J. Mount, J. Ruppert, W. Welch and A. N. Jain, *J. Med. Chem.*, 1999, **42**, 60.
115 A. N. Jain, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 199.
116 A. M. Ghuloum, C. R. Sage and A. N. Jain, *J. Med. Chem.*, 1999, **42**, 1739.
117 W. D. Ihlenfeldt and J. Gasteiger, *J. Comput. Chem.*, 1994, **15**, 793.
118 L. M. Kauvar, D. L. Higgins, H. O. Villar, J. R. Sportsman, A. Engqvist-Goldstein, R. Bukar, K. E. Bauer, H. Dilley and D. M. Rocke, *Chem. Biol.*, 1995, **2**, 107.
119 H. Briem and I. D. Kuntz, *J. Med. Chem.*, 1996, **39**, 3401.
120 U. F. Lessel and H. Briem, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 246.
121 S. L. Dixon and H. O. Villar, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 1192.
122 H. Briem and U. Lessel, *Perspect. Drug Discovery Design*, 2000, **20**, 231.
123 L. J. Soltzberg and C. L. Wilkins, *J. Am. Chem. Soc.*, 1977, **99**, 439.
124 J. H. Schuur, P. Selzer and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 334.
125 C. M. R. Ginn, D. B. Turner, P. Willett, A. M. Ferguson and T. W. Heritage, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 23.
126 V. Schoonjans, F. Questier, Q. Guo, Y. Van der Heyden and D. L. Massart, *J. Pharm. Biomed. Anal.*, 2001, **24**, 613.
127 A. Breindl, B. Beck, T. Clark and R. C. Glen, *J. Mol. Model.*, 1997, **3**, 142.
128 P. Fontana and E. Pretsch, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 614.
129 J. Meiler, W. Maier, M. Will and R. Meusinger, *J. Magn. Reson.*, 2002, **157**, 242.
130 K. Baumann and J. T. Clerc, *Anal. Chim. Acta*, 1997, **348**, 327.
131 H. van de Waterbeemd and E. Gifford, *Nat. Rev. Drug Discov.*, 2003, **2**, 192.
132 T. J. Hou and X. J. Xu, *J. Mol. Model.*, 2002, **8**, 337.
133 W. B. Floriano, N. Vaidehi, G. Zamanakos and W. A. Goddard, 3rd, *J. Med. Chem.*, 2004, **47**, 56.
134 H. Kubinyi, in *Hydrogen Bonding: The Last Mystery in Drug Design?* ed. B. Testa, H. van de Waterbeemd, G. Folkers and R. Guy, Wiley-VCH, Zurich, 2001.
135 R. A. Hill, D. N. Kirk, H. L. J. Makin and G. M. Murphy, in *Dictionary of Steroids*, Chapman & Hill, London, 1991.
136 N. Brooijmans and I. D. Kuntz, *Annu. Rev. Biophys. Biomol. Struct.*, 2003, **32**, 335.
137 A. M. Davis and S. J. Teague, *Angew. Chem., Int. Edn. Engl.*, 1999, **38**, 737.
138 B. P. Morgan, J. M. Scholtz, M. D. Ballinger, I. D. Zipkin and P. A. Bartlett, *J. Am. Chem. Soc.*, 1991, **113**, 297.
139 P. L. Chau and P. M. Dean, *J. Comput.-Aided Mol. Des.*, 1994, **8**, 513.
140 G. Jones, P. Willett and R. C. Glen, *J. Mol. Biol.*, 1995, **245**, 43.
141 G. Klebe and U. Abraham, *J. Med. Chem.*, 1993, **36**, 70.
142 G. E. Kellogg, S. F. Semus and D. J. Abraham, *J. Comput.-Aided Mol. Des.*, 1991, **5**, 545.
143 G. E. Kellogg and D. J. Abraham, *Eur. J. Med. Chem.*, 2000, **35**, 651.
144 I. Pajeva and M. Wiese, *J. Med. Chem.*, 1998, **41**, 1815.
145 P. A. Carrupt, B. Testa and P. Gaillard, in *Reviews in Computational Chemistry*, ed. D. B. Boyd and K. B. Lipkowitz, Wiley, New York, 1997.
146 H.-J. Bohm and G. Klebe, *Angew. Chem., Int. Edn. Engl.*, 1996, **35**, 2588.
147 H. Gohlke and G. Klebe, *J. Med. Chem.*, 2002, **45**, 4153.
148 A. R. Ortiz, M. T. Pisabarro, F. Gago and R. C. Wade, *J. Med. Chem.*, 1995, **38**, 2681.
149 P. Constans and J. D. Hirst, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 452.
150 S. Ren, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1679.
151 M. Goel, D. Jain, K. J. Kaur, R. Kenoth, B. G. Maiya, M. J. Swamy and D. M. Salunke, *J. Biol. Chem.*, 2001, **276**, 39277.

152 K. A. Brown, E. E. Howell and J. Kraut, *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 11753.
153 V. Lukacova and S. Balaz, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 2093.
154 G. Jones, P. Willett and R. C. Glen, in *Pharmacophore perception, development and use in drug design*, ed. O. F. Guner, International University Line, La Jolla, 2000.
155 K. M. Andrews and R. D. Cramer, *J. Med. Chem.*, 2000, **43**, 1723.
156 D. M. Hawkins, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1.
157 D. M. Hawkins, S. C. Basak and D. Mills, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 579.
158 D. M. Hawkins, S. C. Basak and X. Shi, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 663.
159 E. A. Coats, *Perspect. Drug Discovery Design*, 1998, **3**, 199.
160 E. S. Istvan, *Am. Heart J.*, 2002, **144**, S27.
161 Almond - Molecular Discovery Ltd. (http://www.mol-discovery.com).
162 M. A. Kastenholz, M. Pastor, G. Cruciani, E. E. Haaksma and T. Fox, *J. Med. Chem.*, 2000, **43**, 3033.
163 F. Fontaine, M. Pastor and F. Sanz, *J. Med. Chem.*, 2004, **47**, 2805.
164 N. Stiefl, G. Bringmann, C. Rummey and K. Baumann, *J. Comput.-Aided Mol. Des.*, 2003, **17**, 347.
165 M. F. Sanner, A. J. Olson and J. C. Spehner, *Biopolymers*, 1996, **38**, 305.
166 J. L. Pascual-Ahuir and E. Silla, *J. Comput. Chem.*, 1990, **11**, 1047.
167 R. J. Zauhar, G. Moyna, L. Tian, Z. Li and W. J. Welsh, *J. Med. Chem.*, 2003, **46**, 5674.
168 J. T. Clerc and A. L. Terkovics, *Anal. Chim. Acta*, 1990, **235**, 93.
169 K. Baumann, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 26.
170 R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1.
171 Z. Deng, C. Chuaqui and J. Singh, *J. Med. Chem.*, 2004, **47**, 337.
172 S. Putta, C. Lemmen, P. Beroza and J. Greene, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1230.
173 *SYBYL*, Version 6.5.3, HQSAR Module, Tripos Inc., St. Louis, Minnesota, USA.
174 S. Putta, J. Eksterowicz, C. Lemmen and R. Stanton, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1623.
175 T. Kotani and K. Higashiura, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 58.
176 A. Y. Meyer and W. G. Richards, *J. Comput.-Aided Mol. Des.*, 1991, **5**, 427.
177 J. S. Mason and D. L. Cheney, *Pac. Symp. Biocomput.*, World Scientific Press, Hackensack, NJ, USA, 1999, p. 456.
178 J. S. Mason and D. L. Cheney, *Pac. Symp. Biocomput.*, 2000, World Scientific Press, Hackensack, NJ, USA, p. 573.
179 N. Rhodes, P. Willett, A. Calvet, J. B. Dunbar and C. Humblet, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 443.
180 A. J. Hopfinger, S. Wang, J. S. Tokarski, B. Q. Jin, M. Albuquerque, P. J. Madhav and C. Duraiswami, *J. Am. Chem. Soc.*, 1997, **119**, 10509.
181 A. J. Hopfinger, A. Reaka, P. Venkatarangan, J. S. Duca and S. Wang, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1151.
182 J. Liu, D. Pan, Y. Tseng and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 2170.
183 S. Ekins, G. Bravi, S. Binkley, J. S. Gillespie, B. J. Ring, J. H. Wikel and S. A. Wrighton, *Pharmacogenetics*, 1999, **9**, 477.
184 M. D. Krasowski, X. Hong, A. J. Hopfinger and N. L. Harrison, *J. Med. Chem.*, 2002, **45**, 3210.
185 X. Hong and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 324.
186 O. A. Santos and A. J. Hopfinger, *Quant. Struct.-Act. Relat.*, 2002, **21**, 369.
187 A. Vedani, H. Briem, M. Dobler, H. Dollinger and D. R. McMasters, *J. Med. Chem.*, 2000, **43**, 4416.
188 A. Vedani and M. Dobler, *J. Med. Chem.*, 2002, **45**, 2139.
189 M. Whittle, P. Willett, W. Klaffke and P. van Noort, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 449.
190 N. Salim, J. Holliday and P. Willett, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 435.
191 J. W. Godden, L. Xue and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 163.
192 J. D. Holliday, N. Salim, M. Whittle and P. Willett, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 819.
193 M. A. Fligner, J. S. Verducci and P. E. Blower, *Technometrics*, 2002, **44**, 110.
194 R. Wang and S. Wang, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1422.
195 M. L. Verdonk, V. Berdini, M. J. Hartshorn, W. T. M. Mooij, C. W. Murray, R. D. Taylor and P. Watson, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 793.
196 C. M. R. Ginn, P. Willett and J. Bradshaw, *Perspect. Drug Discovery Design*, 2000, **20**, 1.
197 P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *J. Med. Chem.*, 1999, **42**, 5100.
198 J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design: An Introduction*, Wiley-VCH, Weinheim, 1999.
199 R. D. King, S. Muggleton, R. A. Lewis and M. J. Sternberg, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 11322.
200 M. J. E. Sternberg and S. H. Muggleton, *QSAR Comb. Sci.*, 2003, **22**, 527.
201 M. von Korff and M. Steger, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1137.
202 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273.
203 V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, 1995.
204 R. Burbidge, M. Trotter, B. Buxton and S. Holden, *Comput. Chem.*, 2001, **26**, 5.
205 V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk and I. V. Pletnev, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 2048.
206 M. W. B. Trotter and S. B. Holden, *QSAR Comb. Sci.*, 2003, **22**, 533.
207 M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta and C. Lemmen, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 667.
208 J. Weston, F. Perez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff and B. Scholkopf, *Bioinformatics*, 2003, **19**, 764.
209 I. Doytchinova, I. Valkova and R. Natcheva, *SAR QSAR Environ. Res.*, 2002, **13**, 227.
210 H. Lanig, W. Utz and P. Gmeiner, *J. Med. Chem.*, 2001, **44**, 1151.
211 A. D. Harpalani, S. W. Snyder, B. Subramanyam, M. J. Egorin and P. S. Callery, *Cancer Res.*, 1993, **53**, 766.
212 R. Ragno, G. R. Marshall, R. Di Santo, R. Costi, S. Massa, R. Rompei and M. Artico, *Bioorg. Med. Chem.*, 2000, **8**, 1423.
213 X. Q. Huang, T. Liu, J. D. Gu, X. M. Luo, R. Y. Ji, Y. Cao, H. Xue, J. T. F. Wong, B. L. Wong, G. Pei, H. L. Jiang and K. X. Chen, *J. Med. Chem.*, 2001, **44**, 1883.
214 J. P. Horwitz, I. Massova, T. E. Wiese, A. J. Wozniak, T. H. Corbett, J. S. Sebolt-Leopold, D. B. Capps and W. R. Leopold, *J. Med. Chem.*, 1993, **36**, 3511.
215 S. Timofei, L. Kurunczi, W. Schmidt and Z. Simon, *SAR QSAR Environ. Res.*, 2002, **13**, 219.