Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME

6 authors, including:

Andreas Bender
University of Cambridge
331 PUBLICATIONS **7,005** CITATIONS

Catrin Hasselgren
AstraZeneca
46 PUBLICATIONS **833** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   safety View project

Project   □□□□□□ View project

Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME

Robert C Glen[1], Andreas Bender[1]*, Catrin H Arnby[2], Lars Carlsson[2], Scott Boyer[2] & James Smith[3]

Addresses
[1]Unilever Centre for Molecular Science Informatics
Department of Chemistry
University of Cambridge
Lensfield Road
Cambridge
CB2 1EW
UK
Email: andreas.bender@cantab.net

[2]Safety Assessment
AstraZeneca R&D Mölndal
Mölndal
SE-431 83
Sweden

[3]Computer-Chemie-Centrum
Universität Erlangen-Nürnberg
Nägelsbachstraße 25
91052 Erlangen
Germany

*To whom correspondence should be addressed

Circular fingerprints – the representation of molecular structures by atom neighborhoods – have been applied to a wide range of applications, such as similarity searching and the prediction of absorption, distribution, metabolism, excretion and toxicity properties. In recent years there has been a surge in applications resulting from the superior performance of circular fingerprints in comparative studies. This feature examines the nature of circular fingerprints as well as their applications, including virtual screening, metabolism prediction and the estimation of pK_a constants.

Keywords ADME, ADME prediction, circular fingerprints, metabolism, metabolism prediction, pK_a prediction, protonation state, similarity searching, virtual screening

Introduction

A cornerstone of computer-aided drug design is the elucidation of structure-property relationships, meaning the correspondence of the features of a molecular structure to its properties. These properties include those that are physicochemical, such as solubility or lipophilicity (most often expressed as logP), and those that are related to activity against a biological target, such as receptor affinity, but also to more complex responses, such as toxicity, depending on the mode of action that is responsible for the property.

Although the rationale for establishing a correspondence between molecular structure and molecular properties is apparent, both measures have only been defined in vague terms. An important issue is how molecular structure can be represented in a format that is suitable for treatment by computer algorithms. The issue of formatting is integral to two particularly important points related to the requirements of molecular 'descriptors'. First, molecular descriptors require algorithmic definitions to be implemented in computer programs. Second, descriptors must capture those molecular features that are capable of predicting a desired property. The selection of physicochemical or biological effects (or endpoints) must be joined with calculable quantities that have relevance to the estimation of the endpoint. This feature will address the issue of how molecular structures can be represented in a format that is amenable to the computerized generation of structure-property relationships.
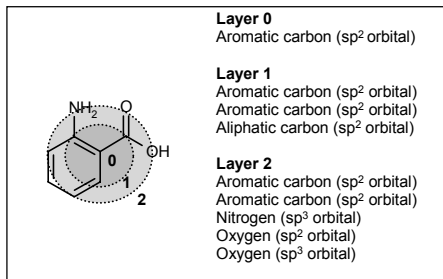
What are 'circular fingerprints'?

Many representations of molecular structures have been devised (reviewed in *Org Biomol Chem* (2004) **2**(22):3204-3218), and these often rely on predefined features. If chosen carefully, these features are able to capture a wide range of information about compounds; however, the features are often biased by preconceived ideas about which structural properties are important and, perhaps more significantly, which ones are not. For cases in which the knowledge of a system is limited, such as an enzyme complex in which only inhibitors are known, but its crystal structure and any clues it could provide for lead optimization are not given, the comprehensive encoding of information is a useful approach.

'Circular fingerprints' follow the route of encoding information, and can capture a large amount of the local structural information that is present in a molecule. They do not, however, capture information on explicit connectivity, which is discussed separately in this feature. The process of generation of circular fingerprints is illustrated in Figure 1. Every heavy atom of a molecule is sequentially used as a starting point for the generation of descriptors and is assigned an atom type. In the example illustrated, the atom type includes information on the elemental atom type and on its hybridization state, as captured by Sybyl mol2 atom types (defined in *J Chem Inf Comput Sci* (2004) **44**(1):170-178). In other circular fingerprints, such as SciTegic fingerprints, different information is encoded for every atom, including its electronic configuration. The atom types for a number of layers, for example, two layers in addition to the central heavy atom, are then assigned to neighboring atoms. To calculate descriptor values, the number of atoms with each given atom type and at each distance from the central atom is recorded. This 'count vector' then serves as a descriptor of the local, chemical environment of the central atom; to create fingerprints for an entire molecule, this process is repeated for each of its atoms. Because the growth of a descriptor is radial, this type of fingerprint is often referred to as a 'radial' or 'circular fingerprint'. An advantage of the approach of recording only the counts of atom types in every given layer is a substantial improvement in computational efficiency, as

there is no need to incorporate a computationally intensive graph-matching approach, which would be necessary for circular fingerprints in which bonding information is also recorded.

Figure 1. Definition of a circular fingerprint: Incorporating information on the arrangement of heavy atoms around each central atom.

> **Layer 0**
> Aromatic carbon (sp^2 orbital)
>
> **Layer 1**
> Aromatic carbon (sp^2 orbital)
> Aromatic carbon (sp^2 orbital)
> Aliphatic carbon (sp^2 orbital)
>
> **Layer 2**
> Aromatic carbon (sp^2 orbital)
> Aromatic carbon (sp^2 orbital)
> Nitrogen (sp^3 orbital)
> Oxygen (sp^2 orbital)
> Oxygen (sp^3 orbital)

This fingerprint is generated for each heavy atom of the molecule. In this example, force field atom types are used that encode the hybridization state of atoms in addition to the elemental atom type; in principle, any type of additional information can be employed.

As with any type of fingerprint, circular fingerprints have certain advantages and disadvantages. While they demonstrated a strong performance in retrospective virtual screening studies (compared in *J Chem Inf Comput Sci* (2004) **44**(5):1708-1718 and *Org Biomol Chem* (2004) **2**(22):3256-3266), metabolism studies (Bayer *et al*: *J Med Chem* (2005): manuscript under preparation) and pK$_a$ prediction studies (*J Chem Inf Comput Sci* (2002) **42**(4):796-805 and *J Chem Inf Comput Sci* (2003) **43**(3):870-879), circular fingerprints omit two significant types of information about molecular structures. First, because the descriptor is local, only a certain number of bonds around a given atom are considered. This factor is partly compensated by the overlap of features, allowing for an implicit connection of disjointed fragments. Second, as only the number of atoms of a given type and their distance is recorded, information about bonding, and the connection of atoms among layers, is not available.

While the local nature of the descriptor and its lack of information on atom bonding limits the information that is stored in circular fingerprints, these types of fingerprints do increase the ability of a model generated for such descriptors to form generalizations from training compounds to new, previously unseen structures. This value is significant, as the 'memorizing the training set' approach is a common feature of many models generated from molecular graphs. From the combination of localized circular fingerprints, different molecular scaffolds can be attributed to the same activity, which increases the ability to perform 'scaffold hopping' (ie, the identification of biologically equivalent structures that have distinct two-dimensional structures) in virtual screening settings.

Applications and comparative studies
Molecular similarity searching/virtual screening
One of the applications of circular fingerprints is molecular similarity searching, or 'virtual screening'. Virtual screening employs a 'molecular similarity principle' to identify

compounds from a database that might show activity against a given molecular target. This identification is achieved by providing one or more 'query' compounds, which are tested in an experimental setting against a target to provide clues as to which molecular features are responsible for the activity. Virtual screening algorithms can currently calculate molecular descriptors for both the query structure and for all library structures. By applying suitable similarity measures, the structures from a database that show similarity to the query structure can be identified, depending on the particular molecular descriptor that is used.

Virtual screening is routinely employed in the pharmaceutical industry, and a number of success stories have been described. To gauge the performance of an algorithm on a new discovery project, its applicability to a series of targets must be proven. Recently, a number of studies on this topic that have evaluated a variety of virtual screening tools on standardized datasets have been published. One of the datasets comprises approximately 1000 structures and five classes of active compounds that span both G-protein-coupled receptor ligands and enzyme inhibitors (serotonin 5-HT$_3$ ligands, angiotensin-converting enzyme inhibitors, hydroxymethylglutaryl coenzyme A reductase inhibitors, platelet-activating factor antagonists and thromboxane A2 antagonists (*Perspect Drug Disc Des* (2000) **20**:231-244). The dataset has been used to evaluate a variety of fingerprint-based as well as structure-based virtual screening methods. A second dataset comprises approximately 102,000 structures and 11 classes of active compounds (*Org Biomol Chem* (2004) **2**(22):3256-3266), including 5-HT$_3$ antagonists, 5-HT$_{1A}$ agonists, 5-HT reuptake inhibitors, dopamine D$_2$ antagonists, renin inhibitors, angiotensin II AT1 antagonists, thrombin inhibitors, substance P inhibitors, HIV protease inhibitors, and inhibitors of cyclooxygenase and protein kinase C. This dataset was also subject to a variety of similarity searching algorithms (discussed in *J Chem Inf Comput Sci* (2004) **44**(5):1708-1718). While both datasets comprise a large number of structures, they are also derived from the MDL Drug Data Repository (MDDR). As such, the datasets have not been tested exhaustively, and many compounds may not be annotated as active simply because they have not yet been tested against a specified target. These datasets are also large because they contain many structural analogs.

From the above studies, circular fingerprints were demonstrated to be the best performing method to generate predictive models when incorporated into a fingerprint on a wide variety of the activity classes detailed above. The relative performance of different methods of fingerprinting is illustrated in Figure 2, which compares the average percentage of active compounds that were retrieved for both circular fingerprint definitions (in black) and other types of fingerprints (in gray).

From the fraction of active compounds that were retrieved from the database containing 102,000 structures, Figure 2 illustrates that the circular fingerprint definitions that were employed, independent of the particular atom type definition, consistently outperformed other fingerprint

Figure 2. Relative performance of a list of fingerprint definitions on a standardized dataset.
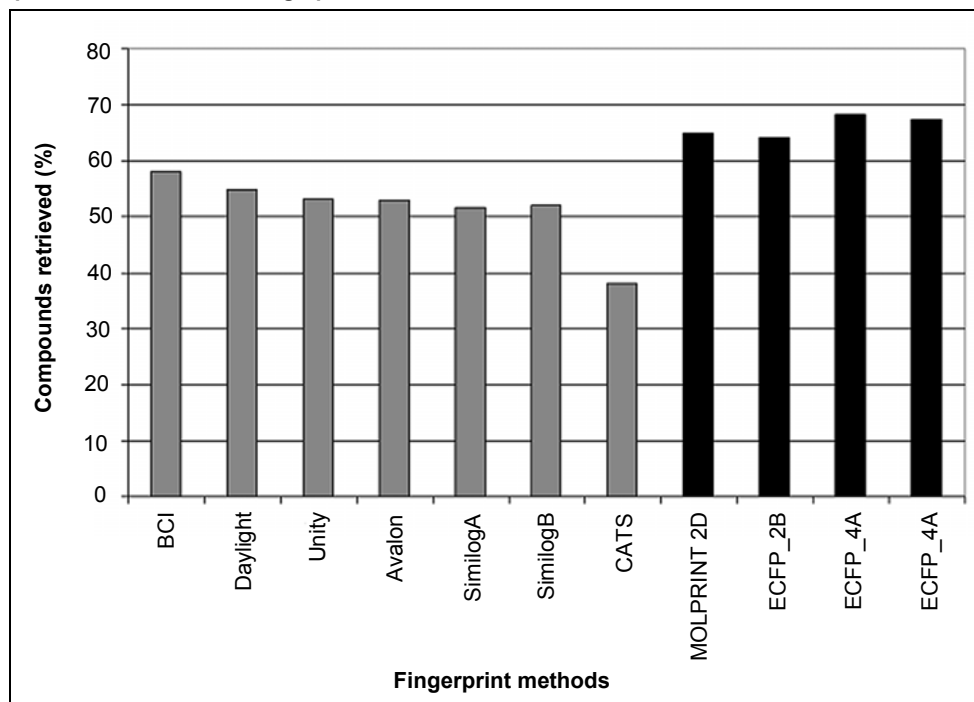
Circular fingerprint definitions are shown in black, while other definitions are shown in gray. Evaluations were conducted in combination with the Bayesian classifier and relevant feature selection. Extended Connectivity FingerPrints (ECFP) are available from SciTegic, while MOLPRINT 2D is defined in *J Chem Inf Comput Sci* (2004) **44**(1):170-178. Barnard Chemical Information (BCI) fingerprints use libraries of fragments, which may include 'fuzzy' definitions; Daylight fingerprints are based on atom paths between two and eight bonds, as well as on augmented atoms; Unity fingerprints employ a mixture of fragments, and the presence of special atoms and atom paths; the Avalon fingerprint is a Novartis AG in-house format, which encodes atoms (both augmented atoms as well as atom triplets); SimilogA and SimilogB are topological three-point pharmacophores; and the CATS descriptor encodes topological two-point pharmacophores. The MOLPRINT 2D method is freely available from *http://www.cheminformatics.org*.

definitions when evaluated in combination with the Bayesian classifier and relevant feature selection. In combination with the Tanimoto similarity coefficient, circular fingerprints also demonstrated improved performance over other fingerprints (*J Chem Inf Comput Sci* (2004) **44**(1):170-178). Circular fingerprints appear to capture information that is relevant to the prediction of bioactivity efficiently and also seem to be suitable descriptors for virtual screening settings. In addition, because of their correspondence to structural features, circular fingerprints can be interpreted unambiguously in many cases (as no bonding information is recorded explicitly, this unambiguous interpretation is not always possible), and can thus provide clues as to which molecular features are responsible for bioactivity. Because circular fingerprints define local regions of a molecular structure, they may be considered as identifiers of the 'patches' on the topology of the molecular structure that are necessary for bioactivity, such as binding.

Prediction of metabolism
Apart from identifying acceptable activity at a desired receptor, which is a challenge that is addressed by concepts such as virtual screening, another of the most important challenges in drug discovery is obtaining favorable pharmacokinetic data of candidate drugs. These data include properties as diverse as the half-life of compounds

and their volume of distribution in the body. The half-life of a xenobiotic in the human body is greatly influenced by its metabolism (reviewed in *J Comput Aided Mol Des* (2002) **16** (5-6):403-413), that is, its biotransformation by enzymes and subsequent elimination from the body. While metabolic enzymes are present in virtually every body tissue, orally administered drugs face hurdles to attain systemic exposure in the body and, indeed, to reach a pharmacological target. Pre-systemic metabolism in the gut involves high concentrations of monooxygenases (eg, cytochrome P450). Following gut metabolism, first-pass metabolic effects in the liver by both monooxygenases and esterases, among many other enzyme families, lead to the formation of metabolites of both drugs and other xenobiotic compounds. The type of metabolism that a compound undergoes is commonly distinguished as phase I metabolism, which adds a polar functionality to the substrate, and phase II metabolism, which adds a polar carrier to the substrate that leads to an increase in solubility and thereby increased elimination.

Structure-based predictions of metabolism have been attempted previously with limited success. The multitude of enzymes involved in metabolic processes means that a multitude of models must be generated, an exercise that is often difficult in practice. A different method based on circular fingerprints was recently developed (Bayer S *et al*: manuscript under preparation). This method relies purely

on biotransformation data, derived from studies of metabolic reactions that occur in a number of species, an approach that circumvents the issue of generating explicit models for each occurring metabolic process. It employs the MDL Metabolite Database as its knowledge base to infer the metabolism of new compounds based on experimental metabolism results reported in the scientific literature. This database currently contains 69,317 transformations compiled from literature sources, with each reaction file containing one or more transformation steps. Rat metabolic data comprise the largest fraction of transformations in the database (47%), followed by human (32%) and dog (9%) data.

The algorithm for establishing the likelihood of transformation at a certain atomic site can be summarized as follows: for a given compound, count the number of substructural environments identified as reaction centers, and count the number of substructural environments identified in the entire database, considering the substrates only. The ratio of these two numbers corresponds to the likelihood that a reaction will occur in this environment, with respect to any kind of transformation reaction. The ratios can then be normalized to provide the relative intramolecular ratios for any given compound.

The method chosen for comparing substructural environments within and between molecules is based upon hierarchical atom type descriptions, or 'six-level circular fingerprints'. These fingerprints are similar to those defined previously, but include a larger number of levels – in this case six as opposed to the two levels used for virtual screening applications. Based on experiments on varying fingerprint dimensions, it appears that metabolic reactions are influenced by the neighboring three to four atoms at the site of interest and therefore require a larger fingerprint depth. For similarity searching applications, descriptors tend to become unique in cases in which a larger number of neighboring layers are considered; unique descriptors for each molecule are not helpful in comparing the similarity of molecules, as they will only identify an exact match. This issue is circumvented by defining a 'fuzzy' matching score. To modify the stringency of the fingerprint matching, two parameters can be varied: the distance and the weighting of the fingerprint level. The general approach involves the requirement for an exact atom match in at least the first two levels of a fingerprint, with a fuzzier match being allowed further away from the primary atom. By allowing for fuzzy (or partial) matching, the knowledge derived from the training set is in a sense 'generalized' to novel cases that show similarity to training set examples, but where no prediction would be possible in exact matching, since no identical matches are present in the database.

Some example metabolic predictions are presented in Table 1. The normalized occurrence ratio for each atom i, represented as $\bar{r}$, is projected as a color onto the structure of a new compound, using RasMol (a molecular visualization software) in this case. The rules for color assignment can be varied to suit the requirements of the user, but in this case the following rules were applied: white, $0 \leq \bar{r}_i < 0.15$; green,

$0.15 \leq \bar{r}_i < 0.33$; orange, $0.33 \leq \bar{r}_i < 0.66$; red, $0.66 \leq \bar{r}_i \leq 1.00$. For readers viewing Table 1 in monochrome, the 'green' sites that are associated with low likelihoods of metabolism are marked as '1', medium sites labeled as '2', and those that are most likely to be metabolized as '3'. To provide an estimate of the prospective performance of the method, those compounds that were present in the MDL Metabolite Database were removed from the dataset, and predictions were confirmed by checking them against their primary literature source. The compounds were evaluated using three different operator settings, which were defined as 'fuzzy', 'standard' and 'strict', depending on similarity requirements for the fingerprint environments to 'match' occurrences in the database. Overall, the metabolic sites for the compounds presented here, as annotated in the MDL Metabolite Database and in the primary literature, could be accurately reproduced.

In the current version of the algorithm, reaction sites are considered if they involve either oxygen addition or bond breakage; in practice this approach comprises important transformation reactions, such as heteroatom oxidations (eg, N-oxidation), aliphatic and aromatic hydroxylations, heteroatom dealkylations (often encountered as N-dealkylations) and amide/ester hydrolysis. (Amide/ester hydrolysis is a result of amidases or esterases and not of cytochrome-P450–mediated reactions.) Because of the nature of this algorithm, which exploits only the knowledge about reactant and product, and not information on mechanistic modeling, it appears to be able to accommodate these different reaction classes.
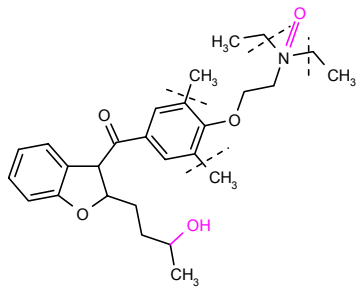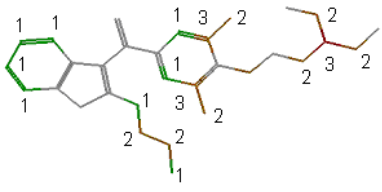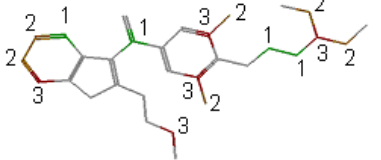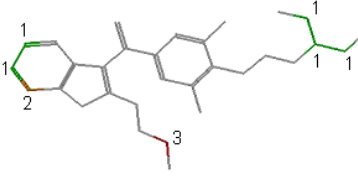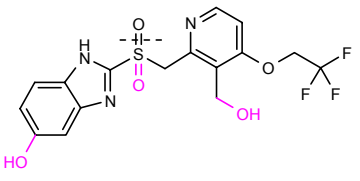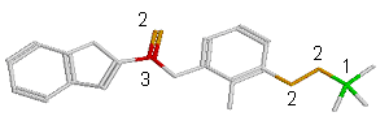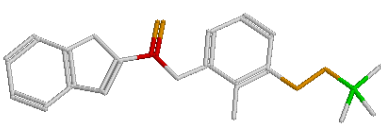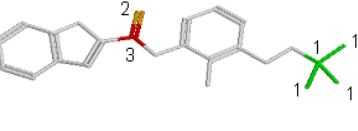
Prediction of pKₐ values
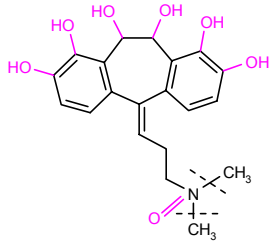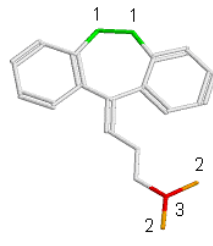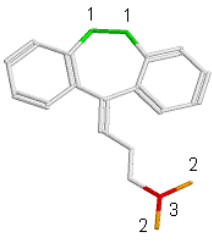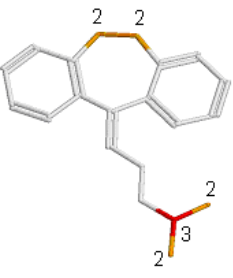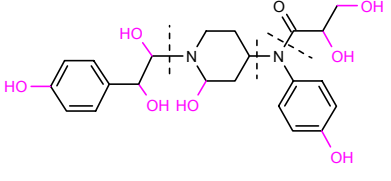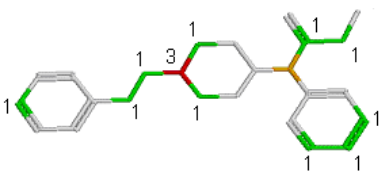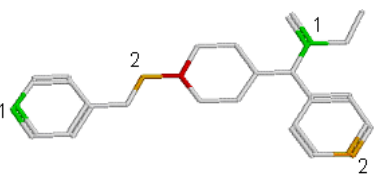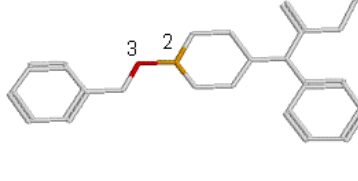
The net electronic charge of a compound at physiological pH, which varies considerably through the intestinal system, is an important determinant of the physicochemical behavior of the compound, as well as its target affinity. The net electronic charge influences the lipophilicity of a compound, and thereby also its solubility, metabolic behavior and virtually every pharmacokinetic property. As shown in the following equation, the protonation state of a compound can be derived from the pKₐ constants of its acidic and basic moieties, which describe the pH at which 50% of the compound is protonated, while the remaining 50% is unprotonated:

$$\% \textit{ ionized} = \frac{100}{1 + 10^{\textit{ (charge(pH - pK}_a\textit{))}}}$$

The application of circular fingerprints to the prediction of pKₐ values is summarized below.

In the first study on the application of circular fingerprints to pKₐ values (*J Chem Inf Comput Sci* (2002) **42**(4):796-805), the prediction of pKₐ values was accompanied by a logP prediction routine based on atom charge and polarizability, and the methods were subsequently applied jointly to the prediction of logD (ie, the protonation-state-adjusted logP) values. In a 10-fold cross-validation on a large dataset (1029

Table 1. Results predicting metabolic sites for selected compounds.

Compound	Experimentally determined sites	Operator settings		
		Fuzzy	**Standard**	**Strict**
Amiodarone				
Amitriptyline				
Fentanyl				
Lansoprazole				

Experimentally determined metabolic sites are listed along with predictions by employing different sets of 'fuzziness' of the descriptor. While slightly different results were retrieved with each setting, the metabolic transforms listed in the literature could be well reproduced in leave-one-out studies, including reactions such as N-dealkylations and hydroxylations, as well as N-oxidations. For the experimental sites, bond breaking is indicated by dashed lines, while all oxygen atoms bound to heteroatoms are results of oxidations (gray bonds and atoms), as are all hydroxyl groups. For the predictions, sites of low metabolism likelihood are indicated as 1, those of medium metabolism likelihood as 2, and those that are most likely to be metabolized as 3.

compounds), q^2 values of 0.85 and 0.83 were obtained for the pK_a predictions of acids and bases, respectively, while the standard error was 0.76 and 0.86 for the models. The pK_a of the majority (> 80%) of the compounds was predicted within 1 log unit of the experimental value, a cutoff that is usually judged to be sufficient in practice. Importantly, in all of the cases observed, the correct rank order of protonations and deprotonations was predicted. In this study, a small number of functional groups were also assigned special group types.

In a follow-up study (*J Chem Inf Comput Sci* (2003) **43**(3):870-879) using a dataset of similar size (1037 compounds), as in the first study, separate, or 'local', models for classes of acids and bases were evaluated. Acids were distinguished among aromatic carboxylic, sulfonic and sulfinic acids; phenols and thiophenols; aliphatic and alicyclic carboxylic, sulfonic and sulfinic acids; aliphatic and alicyclic alcohols and thiols; and acidic nitrogens and carbons. Bases were distinguished among pyridines, anilines, imidazoles and alkylamines. For each of the compound classes, correlation coefficients of > 0.9 (usually > 0.95) were obtained, supporting the rationale behind distinguishing among different compound classes with their different properties.

This second-generation model, as well as the previous model, were then evaluated by applying the models to an external 'standardized' test set consisting of 25 molecules. The new model, consisting of several sub-models, was found to have a significantly higher correlation coefficient on the external test set ($r^2 = 0.99$ versus $r^2 = 0.95$) as well as a lower standard error (0.40 versus 0.70), which supports the rationale behind creating local models for the prediction of pK_a values. Such high predictive values clearly suggest a degree of over-training on the training set.

Summary and outlook

As highlighted in this feature, circular fingerprints provide a molecular representation that performs well in virtual screening, in the prediction of metabolism and in the prediction of pK_a values. Circular fingerprints are easy to calculate and could capture at least as much information as other fingerprinting approaches on the datasets to which they were applied. In addition, their computational efficiency renders them suitable to the processing of very large databases. With current computer technology, a database mining approach of molecular finderprints with 70,000 transformations can be searched in a matter of minutes to make metabolic predictions for new compounds. MOLPRINT 2D can generate and compare molecules at a rate of several thousands per second, which means that a

company collection of 1 million molecules can be searched in approximately 4 min. Computational efficiency is clearly an important factor as database sizes increase.

The application areas of circular fingerprints are currently changing, including a transition from one-target virtual screening approaches to multiple-target classifications. With multiple-target classifications, not only can bioactivity be predicted in the model generated, but the prediction of both desired effects and side effects of substances can also be evaluated. Groups and companies such as PASS and SciTegic, as well as the Unilever Centre for Molecular Science Informatics, are currently performing studies in this direction.

Acknowledgments

Supporting information available
The program MOLPRINT 2D used for generating circular fingerprints from molecular structures, including scripts for performing similarity searches, is freely available from *http://www.cheminformatics.org*.

Further reading

1. Bender A, Glen RC: **Molecular similarity: A key technique in molecular informatics.** *Org Biomol Chem* (2004) **2**(22):3204-3218.

2. Bender A, Mussa HY, Glen RC, Reiling S: **Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier.** *J Chem Inf Comput Sci* (2004) **44**(1):170-178.

3. Bender A, Mussa HY, Glen RC, Reiling S: **Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance.** *J Chem Inf Comput Sci* (2004) **44**(5):1708-1718.

4. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A: **Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures.** *Org Biomol Chem* (2004) **2**(22):3256-3266.

5. Xing L, Glen RC: **Novel methods for the prediction of logP, pK_a, and logD.** *J Chem Inf Comput Sci* (2002) **42**(4):796-805.

6. Xing L, Glen RC, Clark RD: **Predicting pK_a by molecular tree structured fingerprints and PLS.** *J Chem Inf Comput Sci* (2003) **43**(3):870-879.

7. Briem H, Lessel U: *In vitro* and *in silico* **affinity fingerprints: Finding similarities beyond structural classes.** *Perspect Drug Disc Des* (2000) **20**:231-244.

8. Boyer S, Zamora I: **New methods in predictive metabolism.** *J Comput Aided Mol Des* (2002) **16**(5-6):403-413.