CHAPTER 9

Molecular Similarity: Advances in Methods, Applications and Validations in Virtual Screening and QSAR

Andreas Bender,[1,2] Jeremy L. Jenkins,[2] Qingliang Li,[3] Sam E. Adams,[1] Edward O. Cannon[1] and Robert C. Glen[1]

[1]*Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*
[2]*Lead Discovery Center, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Ave., Cambridge, MA 02139, USA*
[3]*College of Chemistry and Molecular Engineering, Center for Theoretical Biology, Peking University, Beijing 100871, China*

Contents

1. INTRODUCTION

Molecular similarity [1–4] follows, in principle, a simple idea: molecules which are similar to each other exhibit similar properties more often than dissimilar pairs of molecules. This is often written as the relationship

$$Property = f(Structure)$$

ANNUAL REPORTS IN COMPUTATIONAL CHEMISTRY, VOLUME 2
ISSN: 1574-1400 DOI 10.1016/S1574-1400(06)02009-3

Which leaves open two major questions:

1. How to represent molecular structure (the connectivity table or the coordinates of atoms are not *per se* suitable choices)?
2. What is the functional form between structure (or rather structural representation) and the property under consideration so that we can derive an empirical measure of similarity?

In order to explicitly include both challenges mentioned one can reformulate to give

$$m(Property) = f\,(g(Structure))$$

where m is the measurement outcome of a molecular property concept (such as log P as a surrogate measure of 'lipophilicity'), g represents the transformation of a molecular structure into a 'descriptor' which is amenable to a statistical analysis or machine-learning treatment and f connects experimental measurement and structural representation. Both steps are generally independent of each other, although some combinations of molecular representation and model generation technique are more sensible than others.

The problem in establishing a suitable function g, which translates a molecular structure into a descriptor representation, is that it is usually not known *a priori* which molecular features contribute to a certain property. For example, some functional groups in ligand–receptor binding will establish ligand–receptor interactions, while others simply point into bulk solvent. Often a large number of descriptors need to be calculated in order to (hopefully) capture the relevant factors for a certain molecular property, since often no direct experimental observation is known.

The problem in establishing a function *f*, which correlates descriptor representation and property is that its functional form is also usually not known. Again, no underlying theory exists and its character can vary between two extremes. Linear regression, for example, represents a simple functional form between input and output variables with the advantage of a very small number of free parameters – and following Occam's razor it should be applied in cases where there is a sound physical reason to believe in an underlying linear relationship between input and output variables. At the other end, neural networks are able to model *any* (also nonlinear) relationships between input and output variables. However, they depend on a large number of variables, which may lead to spurious correlations. Often the choice of a functional form, in the absence of physical laws, is governed simply by trial-and-error.

The problems in establishing the optimal choice of *f* and *g* are increased by the fact that the relationship between structure and measured property (the only relationship available from experimental data!) is rarely given over a large region of chemical space. Data are sparse – estimations of the size of the chemical space for typical drug molecules [5] (up to 30 heavy atoms) are in the region of 10^{60}, experimental datasets on a property of interest are rarely available for more than 10^6 compounds and are often considerably smaller.

A solution to the problem of identifying the 'best' molecular descriptor will never be fully established – for both practical reasons (the limited size of datasets) and theoretical reasons. A wide variety of different features are important for each property and the functional forms between descriptor representation and property can usually not be established from physical laws (and thus cannot be optimized analytically).

Still, we can establish empirical measures of molecular similarity to predict some particular properties better than others, tested on some of the more or less restricted datasets available. This review deals with both novel molecular representations, function g from above, as well as novel model generation and machine learning methods, function f from above.

As soon as a relationship between molecular representation and a particular property's values is established a crucial question arises: how good are predictions for novel molecules?

Ideally, *all* of chemical space would be covered with *zero error*.

Limits in descriptor generation as well as in experimentally available data clearly prevent us from reaching this goal. Still, in order to establish confidence in models in practical settings, this requirement can be replaced by the question:

Which area of chemical space is covered with *acceptable error?*

Different methods (best known among them are approaches like cross-validation), attempt to provide empirical answers to this question. Intuitively one might guess that for the question *which region* is covered by a given model, the distance of compounds from the training set to the novel compounds whose properties are to be predicted is relevant. This is indeed the case, as has been established in recent articles (see Section 4).

The question of how good predictions for novel compounds are is often established by cross-validation, where portions of the available datasets are, in turn, taken as an external test set, while the remainder of the dataset is used for training purposes. The test set thus attempts to simulate a novel set of molecules, unknown to the training phase of the model and root-mean-square errors (RMSE) or cross-validated correlation coefficients (q^2) on the test set are often reported as a measure of the generalizability of models. Recently, it has emerged that cross-validation

actually shows merely that a model is *internally consistent*, but not necessarily predictive for new compounds. The question of how reliability of models can be assured is also discussed in Section 4, and indeed several recent publications propose approaches to determine the 'domain' of models (the area in which they are applicable, see Section 4 for details).

Conventionally, enrichment over random selection is often cited, giving an estimate of how many more active compounds are retrieved from a database than by pure chance. While this measure is correct in the way it is calculated, more recently the performance of 'sophisticated' fingerprints has been compared to trivial features, namely counts of atoms by element, without any structural information [6]. The performance ratio of 'state-of-the-art' methods (i.e., circular fingerprints and Unity fingerprints) to those 'dumb' descriptors can then be interpreted as the 'added value' of more sophisticated methods. Soberingly, on many datasets of actives 'real' fingerprints do not perform significantly better than atom counts (see Fig. 1).

This also relates to the suitability of current databases employed for retrospective virtual screening runs, which are often derived from the MDDR [7,8]. While on the one hand, multiple activity classes are present, those datasets still possess two major disadvantages; first, no information about definite *inactivity* of compounds is contained in the database. Still, if experimental data for retrieved hits are subsequently obtained, many of the 'false-positive' predictions may well be active. Second, following bioisosteric considerations in combination with 'fast follower' approaches to synthesis, it should be noted that this database contains a large number of close analogues. The hit rates obtained on this dataset may thus be overly optimistic compared to real-world libraries employed for virtual screening. Still, the two databases referenced above, which are both subsets of the MDDR, were very important as they enabled comparison of similarity searching approaches on multiple, identical datasets. We would also like to emphasize that more suitable datasets are too often – unfortunately – unavailable from the pharmaceutical and biotechnology companies.

In the following sections, we will also cover other recent developments in some of the areas, which exploit the 'molecular similarity principle'. Section 3 will present novel approaches to capture molecular properties by the use of novel 'descriptors'. Since molecular descriptors and the methods used to analyze the data they represent cannot be separated easily, the second part of this section also covers novel data analysis methods. Section 4 focuses on a crucial aspect of computational models – their validity. In the previous few years, about two dozen publications that focused on 'model validation' have appeared, an area which shall be summarized in this

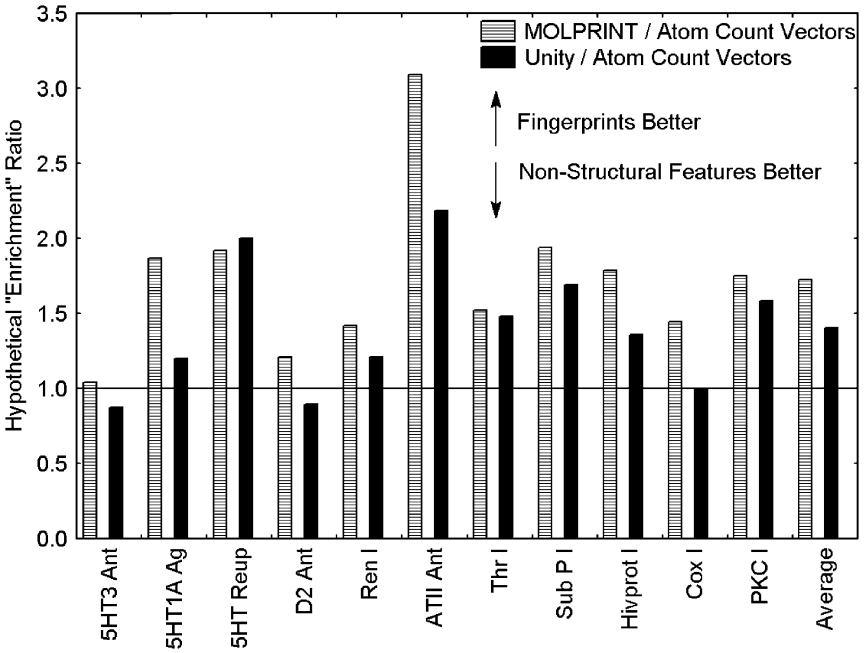

Fig. 1. Comparison of retrieval rates of established descriptors, namely Unity and circular (MOLPRINT 2D) fingerprints, to 'dumb' atom counts. The added value of real descriptors is present in most sets of active compound, although not in all and in many cases only showing low single-digit improvement. Reprinted with permission from J. Chem. Inf. Comput. Sci., 2005, 45, 1372. Copyright 2006 American Chemical Society.

review. Finally, Sections 5 and 6 turn to the application of the methods described earlier. In Section 5, we discuss additional ways to examine data available such as those from high-throughput screening (HTS) campaigns and to gain more knowledge from this data. Section 6 describes some of the recent applications of methods described in the preceding sections, focusing on successes of virtual screening applications, database clustering and comparisons (such as drug- and in-house-likeness) and recent large-scale validations of docking and scoring programs.

2. NOVEL METHODS

We will now describe some of the recent developments in the calculation of molecular descriptors.

2.1. Molecular descriptors

POT-DMC [9] (short for POTency-scaled Dynamic Mapping of Consensus positions) takes not only the (binary) activity of a compound into consideration for virtual screening applications, but also the quantitative activity of a structure. Accordingly, each bit of the descriptor vector (which consists of a combination of one-, two- and three-dimensional (1D, 2D and 3D) features) is multiplied depending on the IC_{50} value of the compound. Scaled bits are summed and normalized at each position. Afterward, the descriptor can be used for virtual screening. When applied to a database of CCR5 chemokine receptor antagonists, serotonin receptor agonists and gonadotropin-releasing hormone agonists, the method overall did not retrieve a larger number of structures – but those which were retrieved were, as intended, of higher activity than in cases where no scaling according to activity was applied.

The FEPOPS[10] (Feature Points of PharmacophoreS) descriptor aims to exploit a (relative) advantage of 3D descriptors, the ability to discover novel scaffolds against a given target, based on active sample structures. After generation of tautomers and conformers, k-means clustering of atomic coordinates is performed. Thus, no knowledge about the active conformation of a structure is necessary. Interaction types are assigned to characteristic 'feature points' in a subsequent step, and are again subject to k-medoids clustering to reduce redundant conformer coverage. Cluster representatives can now be used for similarity searching. Validations are presented using both MDDR (Cox-2, HIV-RT and 5HT3A inhibitors and ligands, respectively) and in-house datasets. In addition, it was shown that inhibitors can be identified from a database, based simply on endogenous ligands (for dopamine and retinoic acid).

A completely different path is followed by the LINGO [11] approach, which is based on a textual representation of molecules. Based on the SMILES string of a structure, and without time-consuming conversion and descriptor generation, a molecule is represented by a set of overlapping 'LINGOs', each of which represents a substring of the complete SMILES structure. While being a straightforward concept (in the best possible sense), favorable performance is presented on log P and solubility datasets, where cross-validated RMS errors are 0.61 and 0.89 log units, respectively. The descriptor also shows applicability to bioactivity, where significant discrimination between bioisosteres and random functional groups can be observed.

Reduced graph descriptors have been the subject of interest for a considerable time, and recently further work was performed in this area with success. Earlier comparison algorithms of reduced graphs represent the

graph as a binary fingerprint, sometimes leading to molecules perceived as similar by the algorithm, which are not similar to the eyes of most chemists. This problem was recently addressed [12] by applying 'edit distance' measures to the similarity of compounds – the number of operations needed to transform one reduced graph structure of a molecule into another. Through this emphasis of not only the fragments present in reduced graphs, but also the way in which they are connected, better agreement with the human perception of 'molecular similarity' could be achieved.

Molecular binding can be thought of as being mediated by complementary shapes and matching properties – where, due to solvation and other effects, 'matching' does not only mean complementarity. Accordingly, a 'Shape Fingerprint' method has recently been presented [13] which implements shape similarity measures akin to volume overlap methods, but which, due to the employment of database-derived reference shapes, is several orders of magnitude faster. (Note of course that shape also plays an important role in other areas of science [14].) Employing Gaussian descriptions of molecular shape, about 500 shape comparisons can be performed per second and the resulting shape similarity was shown to be useful in virtual screening applications.

Only some parts of a ligand bound to its target will actually interact with the target, other parts will just be pointing into the bulk solvent. By analyzing the variability of ligands' regions, features which correspond to each of the regions can be inferred – molecular features which are involved in ligand–target interactions will be more highly conserved than those which point into the solvent, due to the stricter requirements imposed on them. The 'Weighted Probe Interaction Energy (WeP) Method' [15] exploits exactly this principle, and can be used to derive ligand-based receptor models. This was applied to the steroid dataset (which is well known from CoMFA studies) a set of dihydrofolate reductase (DHFR) inhibitors as well as hydrophobic chlorinated dibenzofurans. In particular, the DHFR model was able to elucidate interactions relevant to binding which very closely resemble the target-derived model complex.

Previously applied to the calculation of inter-substituent similarities, which might be exploited for the identification of bioisosteric groups [16], the R-group descriptor (RGD) was more recently also the subject of QSAR investigations [17]. The RGD describes the distribution of atomic properties at a distance of n bonds ($n = 1, 2, 3 \ldots$) away from a core that is common to a series of compounds. In combination with partial least squares the descriptor was applied to several datasets for QSAR studies, comprising of benzodiazepin-2-ones active at $GABA_A$, triazines exhibiting anticoccidal activity and a set of tropanes active at serine, dopamine and

norepinephrine transporters. RGDss in combination with PLS showed comparable performance overall to HQSAR and EVA models in a cross-validation study, in some cases outperforming the other QSAR approaches.

Another alignment-free method for the time-efficient generation of QSAR models is Fingal [18] (a short and straightforward acronym for 'Fingerprint Algorithm'). Unlike RGDs, a hashed fingerprint is generated which encodes structural features of the molecule, where distances may be measured either topologically or by employing spatial information between atoms. Applied to D2 ligands, the 2D version of Fingal, in particular, was able to outperform CoMFA- and CoMSIA-based approaches. For estrogen ligands, performance was highly dependent on the structural class of compounds, not only for Fingal but also for models based on CoMFA, HQSAR, FRED/SKEYS (Fast Random Elimination of Descriptors/Substructure Keys) and Dragon descriptors. In subsets such as a pesticide subset, no model was obtained via CoMFA (correlation coefficient of zero), whereas Fingal gave correlation coefficients as high as 0.85 in a cross-validation study.

The GRID force field [19] has been the basis of a number of descriptors developed recently, among the best-known ones being the GRIND descriptor [20]. Some extensions of the descriptor have been presented recently, which include the incorporation of shape [21] into the descriptor. It was recognized that molecular shape is a major factor determining ligand–receptor binding, a property that was previously not emphasized enough by the original GRIND descriptors. This was due to the fact that only maximum products of interactions are incorporated into the descriptor, omitting large lipophilic features which do not contribute significantly to calculated interaction energies with probes, but might still have profound influence on binding through steric effects. Introducing the new 'TIP' probe (which is not a probe in the traditional sense but a measure of curvature of the molecular surface) led to significant improvements in QSAR studies of adenosine receptor antagonists (of the xanthine structural class) and *Plasmodium falciparum* plasmepsin inhibitors being observed. Interestingly, TIP–TIP correlations were also found to be the most significant descriptors in case of A_1 antagonists, showing the importance of the shape descriptor on this class. The second development was the 'anchor-GRIND' approach [22], which focuses on user-defined features to calculate a distribution of interaction points relative to it, thereby incorporating pre-existing biological knowledge about a target. Models are found to be both of better quality and easier to interpret on congeneric series of hepatitis C virus NS3 protease and of acetylcholinesterase inhibitors, as well as more discriminatory between factor Xa inhibitors of both high and low

affinity. A virtual screening methodology also based on the GRID force field was developed recently [23]. This method was validated on a large dataset containing thrombin inhibitors and also showed potential to select suitable replacements for scaffolds typically encountered in the lead optimization stage.

A molecular 'descriptor' which actually does not employ an explicit transformation of the molecular structure into descriptor space was recently presented [24]. It employs a graph kernel description of the structure in combination with support vector machines (SVMs) for regression analysis. The computational burden is alleviated through employing a Morgan index process as well as the definition of a second-order Markov model for random walks on 2D structures. The method was then validated on two mutagenicity datasets. While already exhibiting the ability to capture molecular features responsible for bioactivity (here mutagenicity) in its current form, future developments might include more abstract representations of the molecular scaffold such as some form of reduced graph representation.

While the bioinformatics area has a multitude of methods which can be applied to the analysis on 1D representations of protein sequences and DNA, due to branching and cyclization the case is far more difficult for small molecules. One of the few 1D representations of molecules [25], based on multidimensional scaling of the structure from 3D into 1D space, has more recently been extended to allow for the alignment of multiple structures [26]. Applied to SKC kinase ligands as well as hERG channel blockers, significant improvement in retrieval rates could be observed in a retrospective study if multiple (in this case 10) ligands were used for screening. The concept of Feature Trees was also recently extended to allow for the incorporation of knowledge derived from multiple ligands into a single query [27], and retrospective screening results on ACE inhibitors as well as adrenergic α_{1a} receptor ligands showed considerable improvements over searches using single queries, both in terms of enrichments as well as the diversity of structures identified.

When structures are encoded in a discrete fashion, 'binning' is often employed in order to convert real-valued distance ranges into binary presence/absence features. This approach is followed in, for example, the CATS autocorrelation descriptor in its 3D version (CATS3D) [28]. However, binning borders may introduce artifacts such that feature distances close to each other but on opposites sides of bin borders being perceived to be as different from each other (simply since features do not match) as much more distant features. Accordingly, a related descriptor termed 'SQUID' was recently introduced which incorporates a variable degree of fuzziness [29]. Applied to Cox-2 ligands considerable retrieval improvement was

observed, with best performance at intermediate degrees of fuzziness. Using Cox-2 ligands as well as thrombin inhibitors in combination with graph-based potential pharmacophore point triangles, typed according to interaction types, features responsible for ligand-target binding could be identified [30]. In addition, prospective screening was performed and a benzimidazole identified as a potent Cox-2 inhibitor was experimentally found to be active in a cellular assay with high affinity ($IC_{50} = 200$ nm).

The ultimate descriptor, in the realm of virtual screening, is the response of the biological system. While structure-derived descriptors are quick and (usually) easy to calculate, they are not the final goal – it is the effect that the compound has in a 'real world' setting. Using those biological effects as descriptors, namely percent inhibition values across a range of 92 targets for a number of 1567 molecules, the 'biospectra similarity' (the similarity of effects on the respective targets) was established via hierarchical clustering [31]. It was found that biospectra similarity provides a solid descriptor for forecasting activities of novel compounds and this was validated by removal of some important target classes after which clustering of compounds was overall still very stable. While the response of single targets is already a step toward biology, protein readouts of cell cultures [32] also incorporate cell signaling networks, thus stepping even closer to whole organism systems (of course at the price of increased complexity and cost involved). Also based on biological response data (phenotypic screening) a 'class scoring' technique was recently developed [33], which does not assign binary (hit/non-hit) activities to individual compounds but to classes of compounds instead. This way, more robust assignments are achieved as well as a lower number of false-positive predictions.

2.2. Data analysis and model generation

SVMs have been previously used for distinguishing, for example, between drug- and non-drug-like structures [34] and recently have been applied in virtual screening [35,36]. Using DRAGON descriptors and a modification of the traditional SVM to rank molecules (instead of just classifying them), performance was in this study [35] validated on inhibitors (or ligands) of cyclin-dependent kinase 2, cyclooxygenase 2, factor Xa, phosphodiesterase-5 and of the α_{1a} adrenoceptor. Compared to methods such as Binary Kernel Discrimination in combination with JChem fingerprints the new approach was found to be superior. The ability of lead hopping was also demonstrated recently through the combination of SVMs with 3D pharmacophore fingerprints (defined as SMARTs queries) [36].

There is a trend in the recent cheminformatics literature toward ensemble methods, i.e., methods where multiple models (instead of a single model) are generated and used together (as an ensemble) to make either qualitative or quantitative predictions about new instances. Random Forests [37] are an ensemble of unpruned classification or regression trees created by bootstrapping of the training data and random feature selection during tree induction. Prediction is then made by majority vote or averaging the predictions of the ensemble. On a set of diverse datasets (blood–brain-barrier penetration, estrogen-binding, P-glycoprotein-activity, multidrug-resistance reversal-activity and activity against COX-2 and dopamine receptors) superior results to methods such as decision trees and PLS were reported. More recently, 'Boosting' was applied to the same (and additional) datasets [38], and as a general rule this new method seems to be slightly superior in large regression tasks, whereas Random Forests are claimed to excel in classification problems. Additionally, employing k-nearest neighbor classifiers, SVMs and ridge regression in an ensemble approach [39] gave significant improvement over single classifiers on a 'frequent hitter' dataset.

Most models derived in QSAR studies, for example, ordinary and partial least-squares regression or principal components regression, employ a linear parametric part and a random error part, the latter of which is assumed to show independent random distributions for each descriptor. However, since molecular descriptors never capture 'complete' information about a molecule, this independence assumption is often not valid. Kriging [40] has replaced the independent errors by, for example, Gaussian processes. Applied to a boiling point dataset and compared to other regression methods (ordinary and partial least-squares and principal component regression) improved performance could be observed.

Alongside model generation, feature selection is also an important step in many studies. Since no perfect descriptors of the molecular system are known, often a multitude of descriptors (often several thousands) are calculated and it is hoped that they capture information, which is relevant to the respective classification or regression task.

A comparative study of feature selection methods in drug design appeared recently [41], which compares information gain, mutual information, χ^2-test, odds ratio and the GSS coefficient (named after the authors, Galavotti, Sebastiani and Simi; a simplified version of the χ^2-test) in combination with the Naïve Bayes Classifier as well as SVMs. While SVMs were found overall to perform favorably in higher-dimensional feature spaces (and do not benefit much from feature selection), feature selection is found to be a crucial step for the Bayes Classifier. (Note that this has at the same time been shown empirically in virtual screening experiments

[42,43].) Some of the methods, namely mutual information and genetic programing, have also been evaluated separately for their use in QSAR studies [44] with respect to a dataset which showed some (typical) problems present in the area, such as a very different sizes of 'active' vs. 'inactive' data subsets.

The problem that structure-activity relationships are rarely linear has been addressed previously through the application of nonlinear methods [45,46] such as k-nearest neighbor approaches [47,48]. More recently, k-nn has also been combined with a CoMFA-like approach, termed k-NN MFA, to predict bioactivity of a compound based on its k-nearest neighbors in 'field space' [49]. As discussed by the authors, some of the disadvantages of CoMFA such as alignment problems are retained; nonetheless, multiple models are produced in each run, giving more room for appropriate model selection. Removing limitations of the statistical model is possible using non-parametric models which have recently been used in QSAR studies [50] and were shown to improve results over more conventional regression-type models. Also Bayesian Regularized Networks have been found to be of interest in recent QSAR studies [51–53]. Those networks possess inherent advantages including that they run less risk of being overtrained than non-Bayesian networks (since more complex models are punished by default).

2.3. New properties of old methods

The effect of binary representations of fingerprints has been known for some time, such as combinatorial preferences [54] and size effects [55] (depending on the similarity coefficient used). More recently, another aspect of the binary representation of features in a fingerprint has been analyzed [56]. Integer or real-valued representations of feature vectors were calculated for 12 activity classes and employed CATS2D and CATS3D autocorrelation descriptors as well as Ghose and Crippen fragment descriptors. Afterward, retrospective virtual screening calculations were performed for both the original (quantitative) representations and the binary (presence/absence) fingerprints. Surprisingly, in only 2 out of the 12 cases did significantly different numbers of actives get retrieved (defined as more than 20% difference). In addition, the retrieved actives showed, depending on the activity class, very different overlap, between 0% and 90%, indicating some orthogonality of the same descriptor, differing by its representation (integer/real-values vs. binary format).

Exploiting the 'molecular similarity principle' by not only looking for neighbors of an active compound and assuming they are active (as is

usually done in virtual screening) but also using this knowledge further to improve the model, has recently been exploited in a method called 'Turbo Similarity Searching' [57]. By feeding back information about the nearest neighbors of an active compound into the model generation step, an increased number of active compounds can be retrieved in a subsequent step. This is analogous to the re-use of hot air in turbo chargers in cars, where the output (hot gas, nearest neighbor in this case) is fed back into the loop to improve performance.

3. METHOD VALIDATION

A number of publications have appeared recently focusing on the validation of QSAR models. A wealth of parameters exist here, such as training/test/validation set splits, the dimensionality of descriptors used in relation to the number of degrees of freedom of a model, or the way selection of features is performed.

While it has been recognized for some time that a larger number of descriptors increases the likelihood of chance correlations [58], more recently a discussion of the validity of statistical significance tests, such as the *F* test, has appeared [59] which puts the number of features considered into relation to the significance of a model. This study cautions in agreement with earlier work that one needs to be very careful when judging the statistical significance of correlation models if feature selection is applied – and that statistically 'significant' models can hardly be 'avoided' if too large a variable pool is chosen to select features in the first place.

Since datasets are generally limited in size, a suitable split into training and test set(s) is crucial in order to achieve sufficient training examples on the one hand, and as high as possible a predictivity of the model on the other. Often, leave-one-out cross-validation has been used to judge model performance – where the compound 'left out' was supposed to be a novel compound found for which property predictions had to be made. Unfortunately this is, according to recent studies, not a suitable validation method [60,61]. In the case of leave-one-out cross-validation, where features are selected from a wider range, the tendency exists in every case to select those features which perform best on a particular compound – thus decreasing generalizability of the model. Results were summarized in a simple statement: 'Beware of q^2!', where specifically the cross-validated correlation coefficient of a leave-one-out cross-validation is alluded to. In addition, general guidelines for developing robust QSAR models were developed, namely a high cross-validated correlation coefficient and a regression, which shows slope close to 1 and no significant bias.

Using theoretical considerations as well as empirical evaluations the question of leave-one-out vs. separate test sets was recently considered in detail [62]. Performing repeated cross-validations of both types on a large QSAR dataset, the conclusion was drawn that in the case of smaller datasets, separate test sets are wasteful, but in case of larger datasets (at least large three-digit numbers of data points) it is recommended. This partly contradicts the above conclusion, that separate test sets should always be used. The discrepancy was explained by the fact that in the earlier work only small separate test sets were used (containing 10 compounds), which was not able to provide a sufficiently reliable performance measure.

The finding that cross-validation often overestimates model performance was corroborated in a recent related study [63], in particular, in cases where strong model selection such as variable selection is applied. The main influence on quality overestimation was found to be a (small) dataset size; other factors are the size of the variable pool considered, the object-to-variable-ratio, the variable selection method, and the correlation structure of the underlying data matrix. While in case of conventional stepwise variable selection overconfidence is commonly encountered, as a remedy LASSO (least absolute shrinking and selection operator) selection is proposed, as well as the utilization of ensemble averaging. Both techniques give more reliable estimates of the quality of the developed model. Given that the latter was shown to improve performance in many cases on its own the generation of reliable performance measures is an additional advantage of ensemble techniques.

Overfitting is a problem which describes good model performance on a training set but much worse performance on subsequent data, and thus, mediocre generalizability of the model (the model is not robust). A recent discussion of this problem, with many accessible examples, gives similar guidelines to those above, such as that leave-one-out cross-validation is not sufficient [64]. It also emphasizes the recommendation of multiple training/test set splits even in the case of very large dataset sizes and of performing cross-validation across classes of compounds in the case of close analogues (instead of molecule-by-molecule splits). In order to have some measure of overfitting, the use of 'benchmark models' such as partial least squares is recommended (depending on the particular problem) in order to determine whether there might be simpler models appropriate to the task (indicating that the more complex model overfits the data).

Using a toxicity dataset of phenols against *Tetrahymena pyriformis* [65] the conclusion that q^2 is not a sufficient predictor for the applicability of a QSAR model to unseen compounds is corroborated, and suggests using the RMS error of prediction (RMSEP) instead. This guideline is presented

along with additional important points: that outliers should not necessarily be deleted since this step reduces the chemical space covered by the model, that the number of descriptors in a multivariate model needs to be chosen carefully and finally that an 'appropriate' number of dimensions is required for PLS modeling. In addition, the influence of the number of variables on predictive performance for training and test sets is investigated.

Several recent publications have attempted to investigate what the actual scope of a QSAR model is – and attempted to develop guidelines to assess the applicability of a model to a novel compound whose properties are to be predicted [66,67]. Two measures for applicability are proposed: the similarity of the novel molecule to the nearest molecule in the training set and the number of neighbors of the novel compound within the training set with a similarity greater than a certain cutoff. As expected, molecules with the highest similarity are best predicted, and this was found to be true across datasets as well as across methods. The applicability measures described above can also be used numerically to derive error bars for estimations of how likely the prediction of a specific model is within a certain error threshold. The issue of model validity was also briefly reviewed from a regulatory viewpoint [68]. In a similar vein, a 'classification approach' has been presented for determining the validity of a QSAR model for predicting properties of a novel compound [69]. Focusing on linear models (though the underlying concept is more generally applicable), the predictions made for compounds within the initial training set are differentiated between 'good residuals' and 'bad residuals'. Using three different datasets (an artemisinin dataset as well as two boiling point datasets) machine-learning methods were employed to predict whether a novel compound belongs to the 'good' or 'bad' class of residuals, thereby making predictions as to whether its properties can be predicted – with a success rate of between 73% and 94%. A stepwise approach for determining model applicability [70] considers physicochemical properties, structural properties, a mechanistic understanding of the phenomenon and, if applicable, the reliability of simulated metabolism in a step-by-step manner. With several QSAR datasets, it could be shown that for substances that are well covered by the training set improved predictions can be made for novel compounds, in agreement with the conclusions stated above.

The performance of similarity searching methods varies widely, comprising both target- and ligand-based approaches. While large enrichment factors (often in the hundreds) are reported, the question arises of how much 'added value' more sophisticated methods actually provide, compared to very simple approaches, and where the gain-to-cost ratio actually shows an optimum. A recent study illustrated that simple 'atom count descriptors' (which do not capture any structural knowledge but represent a

molecule by a set of integers which represent the number of atoms of each element) are able to have comparable performance to state-of-the-art fingerprints [6]. Thus, when averaged over multiple target classes, the added value of virtual screening approaches is probably closer to two (compared to trivial descriptors) than in the region of often published double-digit numbers (compared to random selection). It should be added that performance of 'dumb' and more sophisticated descriptors varied widely, between virtually no difference in performance up to high single-digit performance improvements of state-of-the-art fingerprints (which are, with respect to retrieval rate and on a MDDR-dataset, circular fingerprint descriptors).

4. 'GETTING MORE FROM YOUR DATA'

4.1. Analysis of high-throughput screening data

HTS results are notorious for the amount of noise they contain and methods such as multiple screening runs are routinely applied to alleviate the problem. Still, additional experiments are required. An alternative method was recently presented [71] which, applying purely computational methods, is able to predict truly active compounds with improved reliability in screenings where multiple compounds are screened per well. Using Scitegic circular fingerprints [72], similarities between molecules in wells containing compounds predicted as being active (which may be true positives or, often, just noise) are calculated. The compounds most similar to active compounds are more likely to be active themselves; by predicting (across wells) those compounds which are similar to each other and at the same time are located in wells showing activity, the active compounds out of the mixtures can be estimated. This way, between 29% and 41% of the active compounds could be retrieved in the top 10% of the sorted compounds.

Another approach which attempts to improve knowledge derived from HTS campaigns was recently proposed [73]; the conventional selection of a fixed number of compounds showing activity in a primary screen is replaced for secondary screens ('Top X approach'). Alternatively, methods based on partitioning are frequently employed. In the approach presented here, an ontology-based pattern identification method is employed, which originated from bioinformatics methods (the prediction of gene function based on microarray data). Taking scaffold diversity into account and also applying the 'molecular similarity principle', the overall probability of selecting active compounds from different clusters is maximized. Based on earlier HTS data, significant improvement of hit confirmation rates was

demonstrated, compared to a conventional 'Top X' approach. Related work was recently also performed with a focus on scaffold clustering [74].

As discussed below, scoring functions are not yet able to predict binding affinities sufficiently well across the board of target proteins. Still, the identification of active ligands was shown to be improved by a second data post-processing step. First, ligands are docked to the target. Subsequently, predicted active and inactive compounds are subject to model generation via a Naïve Bayesian model [75] based on circular fingerprints. Applied to protein kinase B and protein-tyrosine phosphatase 1B, significant performance improvements could be observed in combination with Dock, FlexX as well as Glide scores on protein-tyrosine phosphatase 1B. On the other hand, results on protein kinase B results were not improved, which was attributed to the fact that the predicted actives used to train the model were 100% false positives. Understandably, performance cannot be improved if the initial enrichments are not able to identify true positive binders. More recently, another step was introduced between scoring and selecting active and inactive compounds for training the Bayes Classifier [76], which is one of the available consensus scoring methods. Since consensus scoring is often able to rescue docking results in cases where a specific scoring function fails, rank-by-median consensus scoring was shown to improve results for protein kinase B considerably. Other consensus approaches (rank-by-mean, and rank-by-vote) did not perform as well. This was attributed to their sensitivity to cases where one of the scoring functions performs badly. (The median of a set of numbers is less sensitive to outliers than its mean.)

An alternative method for post-processing docking scores is the PostDOCK approach whose final goal is the elimination of false-positive predictions and their discrimination from artifacts [77]. Based on a ligand-target database, derived descriptors (DOCK score, empirical scoring and buried solvent accessible surface area) and models from machine-learning methods were derived to identify false-positive predictions. Validating the method on 44 structurally diverse targets (plus the same number of decoy complexes), 39 of 44 binding and only 2 of 44 complexes were predicted to be of true-positive nature. Compared to purely docking-based methods, DOCK and ChemScore achieve enrichments on the order of five to seven, depending upon the database used, while the method presented here claims to obtain about 19-fold enrichment.

4.2. Consensus predictions

Consensus prediction of docking scores is often able to improve results over single functions and multiple ways have been proposed to combine

scores from different functions such as rank-by-rank, rank-by-vote or rank-by-number [78]. Performance improvement could not be observed in every case and a theoretical study [79] to elucidate the way in which consensus scoring improves results, concluded that this was due to the simple reason that multiple samplings of a distribution are closer to its true mean than single samplings. Assumptions made by the study, such as the performance of each individual scoring function is comparable, have led to the work later being criticized [80], and it has been concluded that consensus scoring *can* improve results but that it is not true in every case (as observed in practice). More recently, it was demonstrated [81] that two criteria are important if consensus scoring is to be successful: first, each individual scoring function has to be of high quality, and second, the scoring functions need to be distinctive. Even if no training data are available to judge those points, rank-vs.-score plots were proposed to gauge the success of target-based virtual screening against a particular target.

While consensus predictions for ligand-based virtual screening have been known for some time, a more recent study extended the descriptors employed to include structural, 2D pharmacophore and property-based fingerprints as well as BCUT descriptors and 3D pharmacophores [82]. Logistic regression and rank-by-sum consensus approaches were found to be most advantageous due to repeated samplings, better clustering of actives (since multiple sampling will recover more actives than inactives) and agreement of methods to predict actives but less so inactives. In addition, more stable performance across a range of targets was observed.

If multiple active compounds are known in a virtual screening setting, the question arises of how to combine the retrieved lists of individual compounds. Applied to different activity classes from the MDL Drug Data Report as well as the Natural Products Database [83] it was recently found that the rank-by-max method generally outperforms the rank-by-sum method, while concluding that the Tanimoto coefficient is superior to 10 other similarity coefficients considered. As to the applicability of consensus approaches, it is found that more dissimilar activity classes profit more than more homogeneous classes, where best retrieval performance is already obtained using lower numbers of query structures (which are then already able to cover the ‘activity island’ inhabited by the particular class of compounds).

5. APPLICATIONS

5.1. Virtual screening

While many applications of virtual screening tools have appeared in the literature, only some examples can be given here.

A phosphodiesterase-4 (PDE4) inhibitor recently has been optimized through the application of small combinatorial libraries [84]. Affinity was increased by three orders of magnitude by screening only 320 compounds after prioritization by FlexX docking. Following the recent SARS scare, a virtual screening procedure via docking (DOCK program) was able to find inhibitors of SARS coronavirus 3C-like proteinase with binding affinities of $K_i = 61\ \mu M$ out of 40 compounds tested [85]. Virtual screening based on a homology model of the neurokinin-1 (NK1) receptor led to the discovery of submicromolar ligands [86], while even nanomolar binding compounds against Checkpoint kinase 1 (CHK1) could be discovered [87] by applying successive filtering for physicochemical properties, pharmacophore filters and docking stages. Ligand-based pharmacophore models generated by Catalyst [88] were used to discover nanomolar ligands of ERG2, emopamil-binding protein (EBP), and the sigma-1 receptor (σ_1) [89]. Out of 11 compounds tested, 3 exhibited affinities of less than 60 nM. High levels of biliary elimination of a CCK2 antagonist led to the quest for novel compounds, which retained activity and selectivity while improving half-life. Using field points derived from XED charges [90], novel heterocycles were proposed [91] (switching from an indole to pyrrole and imidazole series), which decreased molecular weight and polarity and achieved the desired scaffold hop.

Apart from this list of applications against particular targets, only two further applications shall be described here (since the field is simply too large to capture it in its entirety). First, ligand- and target-based approaches were recently compared in their abilities to identify ligands for G-protein coupled receptors [92]. Evaluating docking into homology models, ligand-based pharmacophore models and Feature Trees, 3D similarity searches as well as models built on 2D descriptors, all ligand-based techniques were shown to outperform the docking-based approaches. However, docking also provided significant enrichment.

Second, the 'HTS Data Mining and Docking Competition' presented its results recently [93–95]. Duplicate residual activities of 50,000 compounds against *Escherichia coli* DHFR in primary screening were released in late 2003 [96], upon which 42 groups submitted activity predictions for a test set of the same size (but with unknown activity). Approaches employed ranged from docking [97,98] to purely ligand-based methods [99,100]. Overall, none of them was able to predict actives from the test set reliably. While this was partly due to difference in chemical composition of the training and test sets, an additional problem was posed by the test set which did not contain real 'actives' (showing proper dose-response curves in secondary assays), thus making predictions difficult.

5.2. Clustering

Several novel clustering algorithms have been presented recently, each of which extends previous approaches in its own way. A combination of fingerprint and maximum common substructure (MCS) descriptors [101] speed up clustering (compared to purely MCS methods) enabling its application to large datasets, and the method was shown to be able to identify the most frequent scaffolds in databases, to select analogues of screening hits and to prioritize chemical vendor libraries. A modification of k-means clustering also showed a considerable speed increase to be possible when processing large libraries [102], as demonstrated on a dataset containing about 60,000 compounds derived from the MDDR. The desired speed-up was observed along with favorable enrichment of activity classes within the clusters. By introducing fuzziness into the clustering process [103], superior results can be obtained compared to the original (non-fuzzy or 'crisp') approaches to k-means and Ward clustering, depending on the particular dataset and the property one attempts to predict. Fuzzy clustering assigns partial memberships to multiple classes (instead of binary values); with a log P dataset the best fuzzy parameterization was shown to clearly outperform the best crisp clustering. In addition, partial class memberships were shown to capture the 'chemical character' of a compound more satisfyingly than conventional (crisp) class assignments.

5.3. Drug-likeness and comparison of databases

While the concept of 'drug-likeness' has to be applied with care (and one needs to be aware of its limitations) it has nonetheless received considerable attention in recent years, based on datasets derived from the Available Chemicals Directory (ACD) and the World Drug Index (WDI). First applications employed Ghose/Crippen descriptors in combination with neural networks for classification, and correct classification was achieved for 83% of the ACD and 77% of the WDI, respectively [104]. Later, the application of SVMs was not able to improve overall performance significantly, but the new method was able to correctly classify compounds that were misclassified by the ANN-based technique [34]. Very recently a further analysis of the drug/non-drug dataset appeared, which analyzed SVM performance (as well as that of other machine-learning methods) in more detail [105]. It was found that, in spite of problems with the dataset (some descriptor representations of compounds were, for example, identical in the drug and non-drug dataset) performance could be improved considerably to about 7% misclassified compounds by

optimizing the kernel dimensions employed. An application using 'human-understandable' descriptors of drug- vs. non-drug-like properties has also been presented [106] recently, and was able to distinguish between both datasets with the most important descriptors being proper saturation level and the heteroatom-to-carbon ratio of the molecule. The concept of database comparison is also more generally applicable, as was shown recently when the question of how 'in-house like' external databases are was addressed in order to help to decide whether they should be acquired or not [107].

5.4. Docking validations

A number of validations of docking programs have appeared recently, and it is interesting to observe that they grow in size in every respect – including the number of docking and scoring functions considered as well as the number and diversity of ligand-target complexes employed for their evaluation.

Using DOCK, GOLD and GLIDE in order to evaluate the performance of docking programs in target-based virtual screening on five targets (HIV protease, protein tyrosine phosphatase 1B, thrombin, urokinase plasminogen activator and the human homologue of the mouse double minute 2 oncoprotein), it was concluded that performance is both target- and method-dependent [108]. Performance varied widely, between near-perfect behavior (for example, GOLD in combination with protein tyrosine phosphatase 1b) to negative enrichment (for example, GOLD with HIV protease). Employing FRED, DOCK and Surflex, and adopting the algorithm to the particular binding pocket, it was found that target-based virtual screening is successful in some cases [109], with Surflex probably performing the best overall.

Investigating phosphodiesterase 4B [110] and a set of 19 known inhibitors with 1980 decoys, the scoring functions PMF, JAIN, PLP2, LigScore2 and DockScore were compared with respect to their ability to enrich known ligands. It was found that PMF and JAIN showed high-enrichment factors (greater than four-fold) alone, while a rank-based consensus-scoring scheme employing PMF and JAIN in combination with either DockScore or PLP2 showed more robust results.

In what is probably one of the most extensive studies yet, 14 scoring functions in combination with 800 protein–ligand complexes from the PDBbind database have been compared for evaluation [111]. The scoring functions compared were X-Score and DrugScore, five scoring functions implemented in Sybyl (ChemScore, D-, F- and G-Score and PMF-Score),

four implemented in Cerius2 (LigScore, LUDI, PLP and PMF) as well as two scoring functions implemented in GOLD (GoldScore and ChemScore) as well as the HINT function. Performance was assessed by their ability predicting affinity (K_i/K_d values). Overall, X-Score, DrugScore, Sybyl with ChemScore and Cerius2 with PLP performed better than the other combinations, giving standard deviations in the range of 1.8–2.0 log units.

Another very comprehensive evaluation [112] employed 10 docking programs in combination with 37 scoring functions against eight proteins of seven types. Three criteria were used for assessment, namely the ability to predict binding modes, to predict ligands with high affinity and to correctly rank-order ligands by affinity. While nearly all programs were able to generate crystallographic ligand-target complexes, the identification of the correct structure by the scoring function was found to be considerably more error-prone. Averaged over all targets, none of the programs was able to predict more than 35% of the ligands within an RMSD of equal to or less than 2 Å. While active compounds were correctly identified, activity prediction was more difficult – to the extent that 'for the eight proteins of seven evolutionarily diverse target types studied in this evaluation, no statistically significant relationship existed between docking scores and ligand affinity' [112]. Similar results were obtained on five datasets (serine, aspartic and metalloproteinases, sugar-binding proteins and a 'miscellaneous' set) using the scoring functions Bleep, PMF, GOLD and ChemScore [113], where across all complexes on average no function returned a better correlation than $r^2 = 0.32$.

Interestingly, another recent study drew quite different conclusions from similar observations [114]. Docking endogenous ligands into a panel of proteins it was concluded that proteins are often very promiscuous and do not interact with only a single clearly defined small molecule. While this is surely possible, given the limitations of today's scoring functions it might well be the case that predictions are just not yet good enough.

6. CONCLUSIONS AND OUTLOOK

While a great number of descriptors and modeling methods has been proposed until today, the recent trend toward proper model validation is very much appreciated. Applications of the 'Molecular Similarity Principle' do not yet show the power one would like them to have – and although some of their limitations are surely due to underlying principles and limitations of fundamental concepts, others will certainly be eliminated in the future.

REFERENCES

[1] M. A. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley, New York, 1990.

[2] P. Willett, J. M. Barnard and G. M. Downs, Chemical similarity searching, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.

[3] N. Nikolova and J. Jaworska, Approaches to measure chemical similarity – a review, *QSAR Comb. Sci.*, 2004, **22**, 1006–1026.

[4] A. Bender and R. C. Glen, Molecular similarity: a key technique in molecular informatics, *Org. Biomol. Chem.*, 2004, **2**, 3204–3218.

[5] R. S. Bohacek, C. McMartin and W. C. Guida, The art and practice of structure-based drug design: a molecular modeling perspective, *Med. Res. Rev.*, 1996, **16**, 3–50.

[6] A. Bender and R. C. Glen, A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication, *J. Chem. Inf. Model.*, 2005, **45**, 1369–1375.

[7] H. Briem and U. Lessel, *In vitro* and *in silico* affinity fingerprints: finding similarities beyond structural classes, *Perspect. Drug Discov. Des.*, 2000, **20**, 231–244.

[8] J. Hert, P. Willett and D. J. Wilton, Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1177–1185.

[9] J. W. Godden, F. L. Stahura and J. Bajorath, POT-DMC: a virtual screening method for the identification of potent hits, *J. Med. Chem.*, 2004, **47**, 5608–5611.

[10] J. L. Jenkins, M. Glick and J. W. Davies, A 3D similarity method for scaffold hopping from the known drugs or natural ligands to new chemotypes, *J. Med. Chem.*, 2004, **47**, 6144–6159.

[11] D. Vidal, M. Thormann and M. Pons, LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities, *J. Chem. Inf. Model.*, 2005, **45**, 386–393.

[12] G. Harper, G. S. Bravi, S. D. Pickett, J. Hussain and D. V. S. Green, The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 2145–2156.

[13] J. A. Haigh, B. T. Pickup, J. A. Grant and A. Nicholls, Small molecule shape-fingerprints, *J. Chem. Inf. Model.*, 2005, **45**, 673–684.

[14] J. Wu, R. Tillett, N. McFarlane, X. Ju, J. P. Siebert and P. Schofield, Extracting the three-dimensional shape of live pigs using stereo photogrammetry, *Comput. Electron. Agricult.*, 2004, **44**, 203–222.

[15] C. H. Chae, S. E. Yoo and W. Shin, Novel receptor surface approach for 3D-QSAR: the weighted probe interaction energy method, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1774–1787.

[16] J. D. Holliday, S. P. Jelfs, P. Willett and P. Gedeck, Calculation of intersubstituent similarity using R-group descriptors, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 406–411.

[17] L. Hirons, J. D. Holliday, S. P. Jelfs, P. Willett and P. Gedeck, Use of the R-group descriptor for alignment-free QSAR, *QSAR Comb. Sci.*, 2005, **24**, 611–619.

[18] N. Brown, B. McKay and J. Gasteiger, Fingal: a novel approach to geometric fingerprinting and a comparative study of its application to 3D-QSAR modelling, *QSAR Comb. Sci.*, 2005, **24**, 480–484.

[19] P. J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, *J. Med. Chem.*, 1985, **28**, 849–857.

[20] M. Pastor, G. Cruciani, I. McLay, S. Pickett and S. Clementi, GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors, *J. Med. Chem.*, 2000, **43**, 3233–3243.

[21] F. Fontaine, M. Pastor and F. Sanz, Incorporating molecular shape into the alignment-free Grid-Independent Descriptors, *J. Med. Chem.*, 2004, **47**, 2805–2815.
[22] F. Fontaine, M. Pastor, I. Zamora and F. Sanz, Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent Descriptors, *J. Med. Chem.*, 2005, **48**, 2687–2694.
[23] M. M. Ahlstrom, M. Ridderstrom, K. Luthman and I. Zamora, Virtual screening and scaffold hopping based on GRID molecular interaction fields, *J. Chem. Inf. Model.*, 2005, **45**, 1313–1323.
[24] P. Mahe, N. Ueda, T. Akutsu, J. L. Perret and J. P. Vert, Graph kernels for molecular structure-activity relationship analysis with support vector machines, *J. Chem. Inf. Model.*, 2005, **45**, 939–951.
[25] S. L. Dixon and K. M. Merz, Jr., One-dimensional molecular representations and similarity calculations: methodology and validation, *J. Med. Chem.*, 2001, **44**, 3795–3809.
[26] N. Wang, R. K. Delisle and D. J. Diller, Fast small molecule similarity searching with multiple alignment profiles of molecules represented in one-dimension, *J. Med. Chem.*, 2005, **48**, 6980–6990.
[27] G. Hessler, M. Zimmermann, H. Matter, A. Evers, T. Naumann, T. Lengauer and M. Rarey, Multiple-ligand-based virtual screening: methods and applications of the MTree approach, *J. Med. Chem.*, 2005, **48**, 6575–6584.
[28] U. Fechner, L. Franke, S. Renner, P. Schneider and G. Schneider, Comparison of correlation vector methods for ligand-based similarity searching, *J. Comput.-Aided Mol. Des.*, 2003, **17**, 687–698.
[29] S. Renner and G. Schneider, Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening, *J. Med. Chem.*, 2004, **47**, 4653–4664.
[30] L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider and G. Schneider, Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors, *J. Med. Chem.*, 2005, **48**, 6997–7004.
[31] A. F. Fliri, W. T. Loging, P. F. Thadeio and R. A. Volkmann, Biospectra analysis: model proteome characterizations for linking molecular structure and biological response, *J. Med. Chem.*, 2005, **48**, 6918–6925.
[32] E. C. Butcher, Can cell systems biology rescue drug discovery?, *Nat. Rev. Drug Discov.*, 2005, **4**, 461–467.
[33] J. Klekota, E. Brauner and S. L. Schreiber, Identifying biologically active compound classes using phenotypic screening data and sampling statistics, *J. Chem. Inf. Model.*, 2005, **45**, 1824–1836.
[34] E. Byvatov, U. Fechner, J. Sadowski and G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1882–1889.
[35] R. N. Jorissen and M. K. Gilson, Virtual screening of molecular databases using a support vector machine, *J. Chem. Inf. Model.*, 2005, **45**, 549–561.
[36] J. C. Saeh, P. D. Lyne, B. K. Takasaki and D. A. Cosgrove, Lead hopping using SVM and 3D pharmacophore fingerprints, *J. Chem. Inf. Model.*, 2005, **45**, 1122–1133.
[37] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947–1958.
[38] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan and Q. Song, Boosting: an ensemble learning tool for compound classification and QSAR modeling, *J. Chem. Inf. Model.*, 2005, **45**, 786–799.

[39] C. Merkwirth, H. A. Mauser, T. Schulz-Gasch, O. Roche, M. Stahl and T. Lengauer, Ensemble methods for classification in cheminformatics, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1971–1978.
[40] K. T. Fang, H. Yin and Y. Z. Liang, New approach by Kriging models to problems in QSAR, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 2106–2113.
[41] Y. Liu, A comparative study on feature selection methods for drug discovery, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1823–1828.
[42] A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 170–178.
[43] A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1708–1718.
[44] V. Venkatraman, A. R. Dalby and Z. R. Yang, Evaluation of mutual information and genetic programming for feature selection in QSAR, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1686–1692.
[45] P. Tino, I. T. Nabney, B. S. Williams, J. Losel and Y. Sun, Nonlinear prediction of quantitative structure-activity relationships, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1647–1653.
[46] J. D. Hirst, Nonlinear quantitative structure-activity relationship for the inhibition of dihydrofolate reductase by pyrimidines, *J. Med. Chem.*, 1996, **39**, 3526–3532.
[47] W. F. Zheng and A. Tropsha, Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 185–194.
[48] M. Shen, Y. D. Xiao, A. Golbraikh, V. K. Gombar and A. Tropsha, Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates, *J. Med. Chem.*, 2003, **46**, 3013–3020.
[49] S. Ajmani, K. Jadhav and S. A. Kulkarni, Three-dimensional QSAR using the k-nearest neighbor method and its interpretation, *J. Chem. Inf. Model.*, 2006, **46**, 24–31.
[50] J. D. Hirst, T. J. McNeany, T. Howe and L. Whitehead, Application of non-parametric regression to quantitative structure-activity relationships, *Bioorg. Med. Chem.*, 2002, **10**, 1037–1041.
[51] F. R. Burden and D. A. Winkler, Robust QSAR models using Bayesian regularized neural networks, *J. Med. Chem.*, 1999, **42**, 3183–3187.
[52] F. R. Burden and D. A. Winkler, Predictive Bayesian neural network models of MHC class II peptide binding, *J. Mol. Graph. Model.*, 2005, **23**, 481–489.
[53] T. H. Wang, Y. Li, S. L. Yang and L. Yang, An *in silico* approach for screening flavonoids as P-glycoprotein inhibitors based on a Bayesian-regularized neural network, *J. Comput.-Aided Mol. Des.*, 2005, **19**, 137–147.
[54] J. W. Godden, L. Xue and J. Bajorath, Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 163–166.
[55] J. D. Holliday, N. Salim, M. Whittle and P. Willett, Analysis and display of the size dependence of chemical similarity coefficients, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 819–828.
[56] U. Fechner, J. Paetz and G. Schneider, Comparison of three holographic fingerprint descriptors and their binary counterparts, *QSAR Comb. Sci.*, 2005, **24**, 961–967.
[57] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information, *J. Med. Chem.*, 2005, **48**, 7049–7054.
[58] J. G. Topliss and R. P. Edwards, Chance factors in studies of quantitative structure-activity relationships, *J. Med. Chem.*, 1979, **22**, 1238–1244.

[59] D. J. Livingstone and D. W. Salt, Judging the significance of multiple linear regression models, *J. Med. Chem.*, 2005, **48**, 661–663.
[60] A. Golbraikh, M. Shen, Z. Y. Xiao, Y. D. Xiao, K. H. Lee and A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models, *J. Comput.-Aided Mol. Des.*, 2003, **17**, 241–253.
[61] A. Golbraikh and A. Tropsha, *Beware of q2! J. Mol. Graph. Model.*, 2002, **20**, 269–276.
[62] D. M. Hawkins, S. C. Basak and D. Mills, Assessing model fit by cross-validation, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 579–586.
[63] K. Baumann, Chance correlation in variable subset regression: influence of the objective function, the selection mechanism, and ensemble averaging, *QSAR Comb. Sci.*, 2005, **24**, 1033–1046.
[64] D. M. Hawkins, The problem of overfitting, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1–12.
[65] A. O. Aptula, N. G. Jeliazkova, T. W. Schultz and M. T. D. Cronin, The better predictive model: high q(2) for the training set or low root mean square error of prediction for the test set?, *QSAR Comb. Sci.*, 2005, **24**, 385–396.
[66] L. He and P. C. Jurs, Assessing the reliability of a QSAR model's predictions, *J. Mol. Graph.*, 2005, **23**, 503–523.
[67] R. P. Sheridan, B. P. Feuston, V. N. Maiorov and S. K. Kearsley, Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1912–1928.
[68] J. D. Walker, L. Carlsen and J. Jaworska, Improving opportunities for regulatory acceptance of QSARs: the importance of model domain, uncertainty, validity and predictability, *QSAR Comb. Sci.*, 2003, **22**, 346–350.
[69] R. Guha and P. C. Jurs, Determining the validity of a QSAR model – a classification approach, *J. Chem. Inf. Model.*, 2005, **45**, 65–73.
[70] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela and O. Mekenyan, A stepwise approach for defining the applicability domain of SAR and QSAR models, *J. Chem. Inf. Model.*, 2005, **45**, 839–849.
[71] M. Glick, A. E. Klon, P. Acklin and J. W. Davies, Prioritization of high throughput screening data of compound mixtures using molecular similarity, *Mol. Phys.*, 2003, **101**, 1325–1328.
[72] SciTegic, Inc., San Diego, CA. http://www.scitegic.com.
[73] S. F. Yan, H. Asatryan, J. Li and Y. Zhou, Novel statistical approach for primary high-throughput screening hit selection, *J. Chem. Inf. Model.*, 2005.
[74] S. J. Wilkens, J. Janes and A. I. Su, HierS: hierarchical scaffold clustering using topological chemical graphs, *J. Med. Chem.*, 2005, **48**, 3182–3193.
[75] A. E. Klon, M. Glick, M. Thoma, P. Acklin and J. W. Davies, Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results, *J. Med. Chem.*, 2004, **47**, 2743–2749.
[76] A. E. Klon, M. Glick and J. W. Davies, Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results, *J. Med. Chem.*, 2004, **47**, 4356–4359.
[77] C. Springer, H. Adalsteinsson, M. M. Young, P. W. Kegelmeyer and D. C. Roe, PostDOCK: a structural, empirical approach to scoring protein ligand complexes, *J. Med. Chem.*, 2005, **48**, 6821–6831.
[78] P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins, *J. Med. Chem.*, 1999, **42**, 5100–5109.
[79] R. Wang and S. Wang, How does consensus scoring work for virtual library screening? An idealized computer experiment, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1422–1426.

[80] M. L. Verdonk, V. Berdini, M. J. Hartshorn, W. T. M. Mooij, C. W. Murray, R. D. Taylor and P. Watson, Virtual screening using protein-ligand docking: avoiding artificial enrichment, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 793–806.
[81] J. M. Yang, Y. F. Chen, T. W. Shen, B. S. Kristal and D. F. Hsu, Consensus scoring criteria for improving enrichment in virtual screening, *J. Chem. Inf. Model.*, 2005, **45**, 1134–1146.
[82] J. C. Baber, W. A. Shirley, Y. Gao and M. Feher, The use of consensus scoring in ligand-based virtual screening, *J. Chem. Inf. Model.*, 2006, **46**, 277–288.
[83] M. Whittle, V. J. Gillet, P. Willett, A. Alex and J. Loesel, Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1840–1848.
[84] M. Krier, J. X. de Araujo-Junior, M. Schmitt, J. Duranton, H. Justiano-Basaran, C. Lugnier, J. J. Bourguignon and D. Rognan, Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor, *J. Med. Chem.*, 2005, **48**, 3816–3822.
[85] Z. Liu, C. Huang, K. Fan, P. Wei, H. Chen, S. Liu, J. Pei, L. Shi, B. Li, K. Yang, Y. Liu and L. Lai, Virtual screening of novel noncovalent inhibitors for SARS-CoV 3C-like proteinase, *J. Chem. Inf. Model.*, 2005, **45**, 10–17.
[86] A. Evers and G. Klebe, Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model, *J. Med. Chem.*, 2004, **47**, 5381–5392.
[87] P. D. Lyne, P. W. Kenny, D. A. Cosgrove, C. Deng, S. Zabludoff, J. J. Wendoloski and S. Ashwell, Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening, *J. Med. Chem.*, 2004, **47**, 1962–1968.
[88] Y. Kurogi and O. F. Guner, Pharmacophore modeling and three-dimensional database searching for drug design using catalyst, *Curr. Med. Chem.*, 2001, **8**, 1035–1055.
[89] C. Laggner, C. Schieferer, B. Fiechtner, G. Poles, R. D. Hoffmann, H. Glossmann, T. Langer and F. F. Moebius, Discovery of high-affinity ligands of sigma1 receptor, ERG2, and emopamil binding protein by pharmacophore modeling and virtual screening, *J. Med. Chem.*, 2005, **48**, 4754–4764.
[90] Cresset BioMolecular Discovery Ltd., Letchworth, UK.
[91] C. M. Low, I. M. Buck, T. Cooke, J. R. Cushnir, S. B. Kalindjian, A. Kotecha, M. J. Pether, N. P. Shankley, J. G. Vinter and L. Wright, Scaffold hopping with molecular field points: identification of a cholecystokinin-2 (CCK(2)) receptor pharmacophore and its use in the design of a prototypical series of pyrrole- and imidazole-Based CCK(2) antagonists, *J. Med. Chem.*, 2005, **48**, 6790–6802.
[92] A. Evers, G. Hessler, H. Matter and T. Klabunde, Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols, *J. Med. Chem.*, 2005, **48**, 5448–5465.
[93] C. N. Parker, McMaster university data-mining and docking competition – computational models on the catwalk, *J. Biomol. Screen*, 2005, **10**, 647–648.
[94] N. H. Elowe, J. E. Blanchard, J. D. Cechetto and E. D. Brown, Experimental screening of dihydrofolate reductase yields a ''test set'' of 50,000 small molecules for a computational data-mining and docking competition, *J. Biomol. Screen*, 2005, **10**, 653–657.
[95] P. T. Lang, I. D. Kuntz, G. M. Maggiora and J. Bajorath, Evaluating the high-throughput screening computations, *J. Biomol. Screen*, 2005, **10**, 649–652.
[96] M. Zolli-Juran, J. D. Cechetto, R. Hartlen, D. M. Daigle and E. D. Brown, High throughput screening identifies novel inhibitors of *E. coli* dihydrofolate reductase that are competitive with dihydrofolate, *Bioorg. Med. Chem. Lett.*, 2003, **13**, 2493–2496.

[97] R. Brenk, J. J. Irwin and B. K. Shoichet, Here be dragons: docking and screening in an uncharted region of chemical space, *J. Biomol. Screen*, 2005, **10**, 667–674.
[98] K. Bernacki, C. Kalyanaraman and M. P. Jacobson, Virtual ligand screening against *E. coli* dihydrofolate reductase: improving docking enrichment using physics-based methods, *J. Biomol. Screen*, 2005, **10**, 675–681.
[99] D. Rogers, R. D. Brown and M. Hahn, Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up, *J. Biomol. Screen*, 2005, **10**, 682–686.
[100] A. Bender, H. Y. Mussa and R. C. Glen, Screening for dihydrofolate reductase inhibitors using MOLPRINT2D, a fast fragment-based method employing the naive Bayesian classifier: limitations of the descriptor and the importance of balanced chemistry in training and test sets, *J. Biomol. Screen*, 2005, **10**, 658–666.
[101] M. Stahl and H. Mauser, Database clustering with a combination of fingerprint and maximum common substructure methods, *J. Chem. Inf. Model.*, 2005, **45**, 542–548.
[102] A. Bocker, S. Derksen, E. Schmidt, A. Teckentrup and G. Schneider, A hierarchical clustering approach for large compound libraries, *J. Chem. Inf. Model.*, 2005, **45**, 807–815.
[103] J. D. Holliday, S. L. Rodgers, P. Willett, M. Y. Chen, M. Mahfouf, K. Lawson and G. Mullier, Clustering files of chemical structures using the fuzzy k-means clustering method, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 894–902.
[104] J. Sadowski and H. Kubinyi, A scoring scheme for discriminating between drugs and nondrugs, *J. Med. Chem.*, 1998, **41**, 3325–3329.
[105] K. R. Muller, G. Ratsch, S. Sonnenburg, S. Mika, M. Grimm and N. Heinrich, Classifying 'drug-likeness' with Kernel-based learning methods, *J. Chem. Inf. Model.*, 2005, **45**, 249–253.
[106] S. Zheng, X. Luo, G. Chen, W. Zhu, J. Shen, K. Chen and H. Jiang, A new rapid and effective chemistry space filter in recognizing a druglike database, *J. Chem. Inf. Model.*, 2005, **45**, 856–862.
[107] S. Muresan and J. Sadowski, ''In-house likeness'': comparison of large compound collections using artificial neural networks, *J. Chem. Inf. Model.*, 2005, **45**, 888–893.
[108] M. D. Cummings, R. L. DesJarlais, A. C. Gibbs, V. Mohan and E. P. Jaeger, Comparison of automated docking programs as virtual screening tools, *J. Med. Chem.*, 2005, **48**, 962–976.
[109] M. A. Miteva, W. H. Lee, M. O. Montes and B. O. Villoutreix, Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex, *J. Med. Chem.*, 2005, **48**, 6012–6022.
[110] C. P. Mpamhanga, B. N. Chen, I. M. McLay, D. L. Ormsby and M. K. Lindvall, Retrospective docking study of PDE4B ligands and an analysis of the behavior of selected scoring functions, *J. Chem. Inf. Model.*, 2005, **45**, 1061–1074.
[111] R. X. Wang, Y. P. Lu, X. L. Fang and S. M. Wang, An extensive test of 14 scoring functions using the PDB bind refined set of 800 protein-ligand complexes, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 2114–2125.
[112] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. K. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff and M. S. Head, A critical assessment of docking programs and scoring functions, *J. Med. Chem.*, 2005, ASAP Article; DOI: 10.1021/jm050362n.
[113] P. M. Marsden, D. Puvanendrampillai, J. B. O. Mitchell and R. C. Glen, Predicting protein–ligand binding affinities: a low scoring game?, *Org. Biomol. Chem.*, 2004, **2**, 3267–3273.
[114] A. Macchiarulo, I. Nobeli and J. M. Thornton, Ligand selectivity and competition between enzymes *in silico*, *Nat. Biotechnol.*, 2004, **22**, 1039–1045.