

# Combined similarity and QSPR virtual screening for guest molecules of $\beta$ -cyclodextrin†

Andreas Steffen,<sup>a</sup> Maximilian Karasz,<sup>c</sup> Carolin Thiele,<sup>b</sup> Thomas Lengauer,<sup>a</sup>  
Andreas Kämper,<sup>a</sup> Gerhard Wenz<sup>\*b</sup> and Joannis Apostolakis<sup>\*c</sup>

Received (in Montpellier, France) 24th May 2007, Accepted 20th June 2007

First published as an Advance Article on the web 25th July 2007

DOI: 10.1039/b707856k

We describe a similarity-based screening approach combined with a quantitative prediction of affinity based on physicochemical descriptors, for the efficient identification of new, high affinity guest molecules of  $\beta$ -cyclodextrin ( $\beta$ -CD). Four known  $\beta$ -CD guest molecules were chosen as query molecules. A subset of the ZINC database with 117 695 molecular entries served as the screening library. For each query the 150 most similar molecules were identified by virtual screening against this library with a graph-based similarity algorithm. Subsequently these molecules were scored by means of a QSPR model. The best-scoring, commercially available molecules were selected for experimental verification (14 in total). Binding free energies were determined by isothermal microcalorimetry (ITC). For three of the four queries, at least one ligand with a higher binding affinity than the corresponding query was found. The approach is a promising high throughput alternative to structure-based virtual screening. While  $\beta$ -CD was chosen as a test case because of its technical relevance and the availability of many binding data, the applied methodology is transferable to other host–guest systems.

## Introduction

The rational design of novel host–guest systems is of particular interest in supramolecular chemistry.<sup>1</sup> Traditionally, their identification has often been based on trial and error, experience and intuition. However, recently studies have been published in which computational methods—borrowed from the field of drug design—were applied to synthetic host–guest systems in order to identify optimally interacting systems. De Jong *et al.*,<sup>2</sup> for example, performed a virtual screening for novel guest molecules of a  $\beta$ -cyclodextrin ( $\beta$ -CD) dimer by means of the protein–ligand docking tool DOCK.<sup>3</sup> About 110 000 substances were virtually screened and ranked. 30 of the manually inspected top-ranking molecules were proposed for further experimental verification. Nine guest molecules showed strong binding affinity with *in vitro* testing. Corbellini *et al.* applied a similar approach to search for guest molecules of a molecular capsule.<sup>4</sup> From about 30 000 virtually screened

substances, a restricted number of the computationally predicted binders were selected for experimental testing, which led to five compounds demonstrating strong encapsulation as tested by NMR spectroscopy. We recently presented an efficient protocol that focused on the design of a  $\beta$ -CD-based receptor for the anticancer drug camptothecin by means of an inverse virtual screening strategy.<sup>5</sup> A virtual library of  $\beta$ -CD derivatives was generated and virtually screened with docking tools. With this approach, we were able to identify six new synthetic receptors for camptothecin, with a binding affinity significantly higher than other receptors from the literature.

These approaches have demonstrated the potential and the possible impact of structure-based virtual screening methods adapted from drug design for the optimisation of synthetic host–guest systems. Similarity- or ligand-based screening is an alternative virtual screening technique that uses information on known guest molecules. This technique relies on the assumption that structurally similar molecules exhibit similar binding properties with respect to a given target. In general, similarity-based screening is faster than structure-based screening. Several approaches have been proposed to describe similarity between molecules. Some rely on the comparison of molecular graph topology,<sup>6</sup> others compare the three-dimensional shape of molecules.<sup>7</sup> A very important and widely applied class of similarity tools employs so-called molecular fingerprints, which are bit-string representations of molecules.<sup>8–10</sup> Similarity-based screening has already been used successfully for the identification of new drugs for given protein receptors.<sup>11</sup>

In our work, a graph-based similarity method was combined with a quantitative structure–property relationship (QSPR) model. This model provides the means to estimate

<sup>a</sup> Max-Planck-Institut für Informatik, Computational Biology and Applied Algorithmics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany. E-mail: asteffen@mpi-sb.mpg.de; Fax: +49 681 9325399; Tel: +49 681 9325328

<sup>b</sup> Organische Makromolekulare Chemie, Saarland University, Geb. C4.2, D-66123 Saarbrücken, Germany. E-mail: g.wenz@mx.uni-saarland.de; Fax: +49 681 302 3909; Tel: +49 681 302 3449

<sup>c</sup> Institute for Informatics, Research and Educational Unit for Bioinformatics and Practical Informatics, Ludwig-Maximilians-University, Amalienstr. 17, D-80333 München, Germany. E-mail: apostola@bio.ifi.lmu.de; Fax: +49 89 2180 4054; Tel: +49 89 2180 4057

† Electronic supplementary information (ESI) available: Details on the calculation of the similarity, selected descriptors of the SVMR model and additional figures of the nested cross-validation. See DOI: 10.1039/b707856k

the binding free energy  $\Delta G^\circ$  of the similarity hits and can be used as a second filter. Compared to docking tools, QSPR models are more efficient and often more accurate within the chemical neighbourhood of the molecules of the training set. On the other hand, a certain amount of experimental binding data is required, whereas in docking the structure of the receptor has to be available. The value of QSPR models for the prediction of  $\Delta G^\circ$  values of complexes between various guest molecules and  $\beta$ -CD was shown in two recent studies.<sup>12,13</sup> In both cases, stable and good predictive models could be generated on the basis of computed molecular descriptors.

$\beta$ -CD is a cyclic oligomer that consists of seven  $\alpha$ -(1–4)-linked D-glucose units.<sup>14,15</sup> Since  $\beta$ -CD is nearly non-toxic and is able to bind small molecules (guests) within its hydrophobic cavity, it is already in use as a solubilizer,<sup>16</sup> e.g. for hydrophobic drugs, leading to several new formulations for drugs already on the market.<sup>17–19</sup> Furthermore,  $\beta$ -CD is used in catalysis,<sup>20</sup> as a protecting agent for unstable molecules, such as vitamins,<sup>21</sup> and in consumer cosmetics.<sup>22</sup> The industrial relevance, together with the availability of binding affinity data, was the motivation for choosing  $\beta$ -CD as the host molecule for our study. However, the proposed protocol can be applied to other host–guest systems of interest. Here we provide a proof of principle for the combined approach as a method for the efficient and reliable identification of novel guest molecules for a given host molecule.

## Methods

Fig. 1 illustrates the workflow of our study. First, a QSPR model for the prediction of  $\Delta G^\circ$  of  $\beta$ -CD inclusion complexes was generated. Second, a similarity-based screening was performed. Then the QSPR model was used for assessing molecules that were found by similarity-based virtual screening. We selected molecules with a predicted high affinity for  $\beta$ -CD in order to experimentally verify our computations.

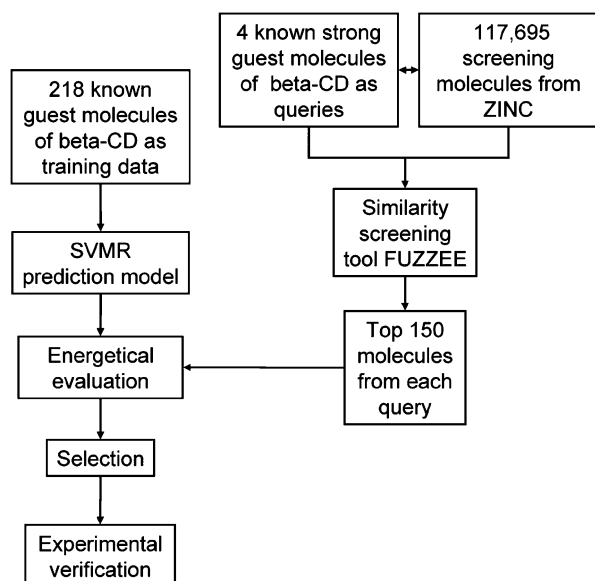


Fig. 1 Schematic flow of the applied virtual screening method.

## Generation of a support vector machine regression based QSPR model

We derived a support vector machine regression (SVMR)-based QSPR model for predicting the binding free energy  $\Delta G^\circ$  of  $\beta$ -CD inclusion complexes based on molecular descriptors<sup>23</sup> and experimental data from the literature. The molecules of our training dataset were taken from Suzuki.<sup>24</sup> All 218 molecules form 1 : 1 inclusion complexes with  $\beta$ -CD. They were drawn with reasonable protonation states with ISIS/Draw<sup>25</sup> and exported as MOL files. CORINA<sup>26</sup> was used to generate low energy 3D structures in the structure data file (SDF) format. Finally, all molecules were visually inspected and, if necessary, corrected.

1666 molecular descriptors were calculated for each molecule by means of the E-Dragon web server.<sup>23,27</sup> The descriptors account for simple molecular properties, from molecular weight and topological features up to elaborate quantum mechanical characteristics. For the subsequent development of the QSPR model, all properties were scaled to the range  $[-1, 1]$  to avoid numerical problems and prevent a bias in the descriptor space. SVMR was chosen for our study because a prior comparison of multiple regression models has shown that SVMR provides particularly stable and predictive regression models for the investigated system and is also relatively simple to apply.<sup>28</sup> In comparison to conventional multiple linear regression we found that SVMR performs slightly better on the given data, however the difference is small. The theoretical background of SVMR has been described in detail by Drucker *et al.*<sup>29</sup> SVMR is an extension of support vector machines (SVMs),<sup>30</sup> which are applied to classification problems. SVMs identify the hyperplane that separates positive from negative examples with a maximum margin. This margin is defined as the distance of the closest data point to the separating hyperplane. The statistical model produced in this manner only depends on a subset of the training data. This subset consists of those data points that are close enough to influence the size of the margin and the orientation of the hyperplane. These are the most difficult examples in the training set and are called the support vectors, since they define the orientation of the separating hyperplane. In SVMR the same effect is achieved by the use of a so-called  $\varepsilon$ -insensitive cost function, which ignores errors up to a defined threshold during the generation of the model. Thus, any training data being predicted by the current model with an accuracy of up to  $\varepsilon$  can be neglected. In this work, we use the LIBSVM implementation<sup>31</sup> with the linear kernel function and combine it with a forward descriptor selection procedure. The latter helps to limit the number of integrated descriptors, which enhances the interpretability of the regression model. Furthermore, the risk of overfitting the model to the underlying data and thereby decreasing the predictability of the model for non-training molecules is reduced if the number of integrated descriptors is limited. The forward descriptor selection procedure is based on a greedy heuristic that proceeds iteratively through a number of steps. First, a regression model is generated for each single descriptor with tenfold cross-validation. Second, the descriptor that gives the highest linear correlation coefficient  $r^2$  is chosen. Then this descriptor is

combined with each of the remaining descriptors and the pair that leads to the regression model with the highest  $r^2$  value is selected for the next descriptor extension step. This is repeated until a maximum for  $r^2$  is reached (see Fig. 4). We stopped the addition of descriptors at this maximum in order to have a well-defined stop criterion. It should, however, be noted that other stop criteria can be defined as well that, e.g., consider the significance of an added descriptor.

The final model was used for predicting the  $\Delta G^\circ$  value of the inclusion complexes between  $\beta$ -CD and the guest molecules that were identified by virtual screening (see below).

**Internal validation of the QSPR model.** The squared linear correlation coefficient  $r^2$ , derived from a tenfold cross-validation test, is known to be overoptimistic with respect to the prediction accuracy of unseen data, especially when cross-validation is used to select the descriptors for the model. A more realistic estimate of the predictability of the final model generated in the described manner can be obtained if a nested cross-validation protocol is performed.<sup>32</sup> Therefore, the data was randomly split into three equally sized subsets  $S_1$ ,  $S_2$  and  $S_3$ . Out of each possible pairing of the three subsets, three combined subsets  $V_1$  ( $S_1 + S_2$ ),  $V_2$  ( $S_1 + S_3$ ) and  $V_3$  ( $S_2 + S_3$ ) were built. Each of the latter served as a training set for the generation of a QSPR model with which the remaining, unused subset was predicted. The prediction quality of the model on these test sets is taken to mirror the applicability to unseen data.

### Virtual screening

Four known  $\beta$ -CD guest molecules **1–4**, with  $\Delta G^\circ$  values less than or equal to  $-20 \text{ kJ mol}^{-1}$ , were selected as query compounds (see Table 1). Two of them, flurbiprofen **3** and ibuprofen **4** are drug molecules. The query compounds were

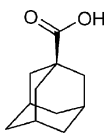
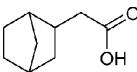
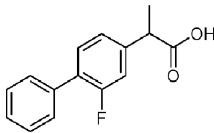
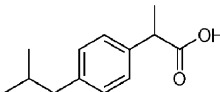
prepared with the same protocol as described for the preparation of the QSPR training set.

The screening dataset was downloaded from ZINC<sup>33</sup> as SDF files. For reasons of direct and fast commercial availability we chose the Sigma Aldrich subset. Altogether this subset contained 117 695 entries. The structures were taken as provided from ZINC (see Irwin and Shoichet<sup>33</sup> for more details of their preparation protocol).

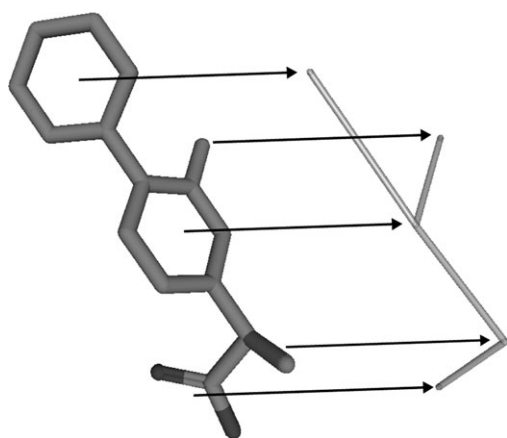
The approach used for similarity screening is based on a variant of the graph matching algorithm used by the GMA program.<sup>34</sup> Direct graph matching at the atomic level leads to the identification of chemically closely related structures. In order to find molecules of different topology, yet similar physicochemical features, it is preferable to perform the comparison on a more abstract representation of the molecules, for example, at the level of functional groups. The computational representation of molecules used in this work is related to the reduced graphs used by Gardiner *et al.*<sup>35</sup> and is illustrated in Fig. 2.

Reduced graphs describe molecules as a collection of connected functional groups or fragments. Each node in the graph represents a fragment in the molecule. Edges between the nodes represent the connectivity of the corresponding fragments. The fragments are obtained as follows: rings containing up to seven atoms form a fragment. Larger rings are fragmented according to the rules for linear chains. Atoms that belong to more than one ring are assigned to each of the respective fragments. Furthermore atoms with at least two non-hydrogen neighbours form the basis of a fragment. The remaining atoms with only one non-hydrogen neighbour are merged into their neighbour's fragment, unless their neighbour is member of a ring fragment. In this case, the atom forms a single atom fragment. Two nodes are connected if they share one or more atoms, or if two of the contained atoms are connected to each other by a chemical bond.

**Table 1** Known  $\beta$ -CD guest molecules that served as query compounds. Experimental error of  $\Delta G^\circ \pm 0.3 \text{ kJ mol}^{-1}$

ID	Name	Structure	CAS-No.	$\Delta G^\circ/\text{kJ mol}^{-1}$	Literature
1	Adamantane-1-carboxylic acid		828-51-3	-24.9	Harrison and Eftink <sup>38</sup>
2	2-(3-Bicyclo[2.2.1]heptyl)acetic acid		1007-01-8	-20.8	Godinez <i>et al.</i> <sup>39</sup>
3	2-[(3-Fluoro-4-phenyl)phenyl]propanoic acid		5104-49-4	-18.8	Ueda and Perrin <sup>40</sup>
4	2-[4-(2-Methylpropyl)phenyl]propanoic acid		15687-27-1	-22.6	Wenz <sup>a</sup>

<sup>a</sup>  $\Delta G^\circ$  value determined within our lab according to the protocol described in section *Binding studies*.



**Fig. 2** Reduced graph representation of flurbiprofen (left: atomic level; right: reduced graph representation). Hydrogen atoms are omitted for clarity.

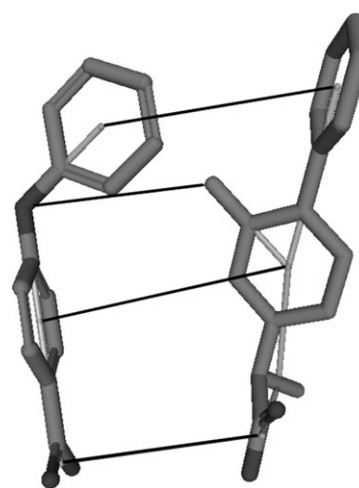
Each node obtains a number of features describing the atoms that constitute the original fragment. The features used are shown in Table 2. Each feature has a weight and a value, which counts the occurrences of the feature in the fragment. The overall similarity between two molecules is the sum of the similarities of mapped nodes. This similarity is then normalised by dividing by the maximum of the self similarities of the two molecules (see ESI for detailed formulas†).

An example of molecule matching is given in Fig. 3, where the matching parts of the molecules flurbiprofen and 4-phenoxybenzoic acid are depicted.

For each of the four query compounds **1–4**, a virtual screening run was performed against the screening dataset. Ranking lists were derived from the calculated similarity scores. The top-ranking 150 molecules of each of the ranking lists were scored by means of the generated QSPR model. The aim of our study was to identify molecules with low  $\Delta G^\circ$  values in complex with  $\beta$ -CD. From the screening runs only those molecules were selected for which a lower or comparable  $\Delta G^\circ$  value with respect to the corresponding query structure was predicted. Furthermore, we were interested in identifying novel molecular scaffolds and thus only molecules with a significant change in the structure compared to the query structure were considered. Additionally, we limited ourselves to commercially available molecules with promising water solubility to allow experimental determination of  $\Delta G^\circ$  by microcalorimetry.

**Table 2** Features of nodes and weighting scheme

Index	Weight	Feature
1	1	Carbon $sp^3$
2	1	Carbon $sp^1/sp^2/ar$
3	1	Nitrogen $sp^3$
4	1	Nitrogen $sp^1/sp^2/ar$
5	1	Oxygen
6	1	Phosphorus
7	1	Sulfur
8	1	Halogens
9	1	Other atom types
10	4	H-Bond donor base
11	4	H-Bond acceptor



**Fig. 3** The matching between flurbiprofen (right) and 4-phenoxybenzoic acid (left). Hydrogen atoms are omitted for clarity.

### Binding studies

Compound **4** was purchased from Avocado, compounds **16** and **9** from Fluka, compounds **5**, **8**, **13**, **15** and **17** from Sigma, compounds **7**, **11**, **14** and **18** from Aldrich and compounds **10** and **12** from Acros Organics.

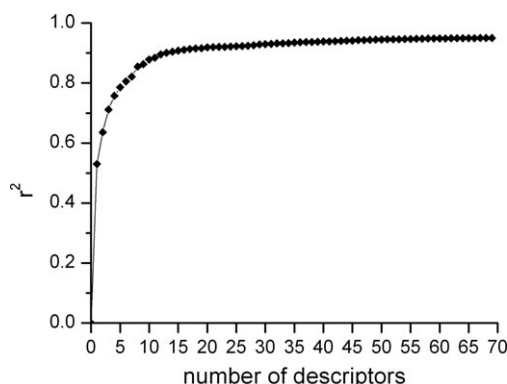
$\Delta G^\circ$  values of the complexes between  $\beta$ -CD and the compounds that were sufficiently water soluble were measured using isothermal microcalorimetric titration at a temperature of 25.0 °C with an AutoITC isothermal titration calorimeter (MicroCal Inc., Northampton, USA) using 1.4144 ml sample and reference cells. The reference cell was filled with distilled water. The sample cell was filled with a 1.3 mM solution of the respective guest in 25 mM phosphate buffer (pH 6.79) and constantly stirred at 450 rpm. A 13 mM solution of  $\beta$ -CD was prepared in the same buffer. This solution was automatically added by a syringe in 20 separate injections of 12.5  $\mu$ l. The resulting 20 heat signals were integrated to yield the mixing heats, which were corrected by the corresponding dilution enthalpies of  $\beta$ -CD. The titration curve was fitted by non-linear regression. Thereby a 1 : 1 stoichiometry of the inclusion compound and the host molecule was appropriate. The binding constant  $K$  and the molar binding enthalpy  $\Delta H^\circ$  were obtained as fitting parameters, from which the binding free energy  $\Delta G^\circ$  and binding entropy  $\Delta S^\circ$  were derived.

### Results and discussion

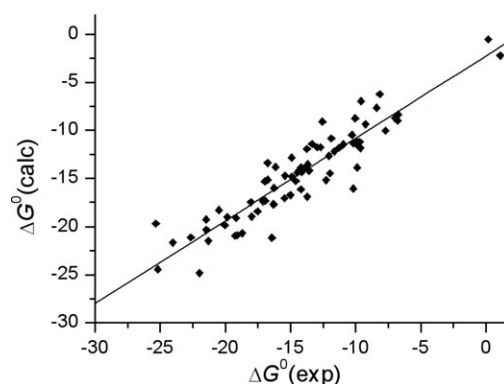
In the first step a QSPR model was generated based on a dataset consisting of 218 molecules.<sup>24</sup> From 1666 molecular descriptors, calculated with E-Dragon, finally 68 descriptors were selected which described the data with an  $r^2$  value of 0.95 and a root mean squared error (RMSE) of 1.17 kJ mol<sup>-1</sup> (Fig. 4). The observed correlation is in good agreement with the one reported by Suzuki.<sup>24</sup> ( $r^2 = 0.92$ ), indicating that the chosen regression methodology and the computed descriptors are appropriate for our study. In contrast to previous descriptors, the E-Dragon descriptors are freely accessible.

For the validation of our approach, a nested cross-validation protocol was used. As described in the Methods section,





**Fig. 4** Descriptor selection for the training set. For 68 descriptors the maximal  $r^2$  value of 0.95 and an RMSE of 1.17 kJ mol<sup>-1</sup> were found.



**Fig. 5** Prediction of  $\Delta G^\circ$  (kJ mol<sup>-1</sup>) for subset  $S_1$  by means of the regression model for validation set  $V_3$ . ( $r^2 = 0.85$ , RMSE 1.98).

for each of the three validation sets models were generated with the same procedure as applied for the entire final model. Each model was then used to predict the  $\Delta G^\circ$  values of the corresponding unused molecules. The mean  $r^2$  value of the three sets is  $0.84 \pm 0.01$ . The RMSE for the three sets is  $2.06 \pm 0.23$  kJ mol<sup>-1</sup>. We note that the corresponding experimental error lies below 0.3 kJ mol<sup>-1</sup>. Reiteration of the splitting and the nested cross-validation process produces very similar results on average. The predicted  $\Delta G^\circ$  values of set  $S_1$  correlated well with the corresponding experimental values, as shown in Fig. 5. The remaining plots can be found in the ESI.†

The 150 most similar molecules were searched for each of the four query structures **1–4** out of the screening dataset of 117 695 entities. All similarity screening runs together took approximately 1 h on a single Xeon 2.8 GHz CPU. This includes preprocessing of the database. If the preprocessing

is precalculated and stored, the similarity screening takes several minutes per compound. From these 4 subsets of 150 top-ranking molecules, the 14 most promising and commercially available molecules were selected by means of the generated QSPR model (Tables 3–6) and visual inspection.

The subsequent experimental testing revealed that only one molecule displayed no binding affinity at all. Ten molecules exhibited a binding free energy of about  $-20.0$  kJ mol<sup>-1</sup> or less. Five of them (**9**, **12**, **14**, **16** and **17**) showed even stronger binding affinities than the corresponding query. Thus, for three of the four screenings, at least one ligand was found with a better affinity than the original query. This is a good result considering that on average only 3–4 new compounds were tested per query. The RMSE of the predicted values to the experimentally determined ones was 2.9 kJ mol<sup>-1</sup>, only slightly higher than the RMSE obtained in the cross-

**Table 3** Selected guest molecules derived from the virtual screening for query **1**

ID	Name	Structure	CAS-No.	Similarity	$\Delta G^\circ$ predicted/ kJ mol <sup>-1</sup>	$\Delta G^\circ$ experimental/ kJ mol <sup>-1</sup>	$\Delta H^\circ$ experimental/ kJ mol <sup>-1</sup>	$T\Delta S^\circ$ experimental/ kJ mol <sup>-1</sup>
1	Adamantane-1-carboxylic acid <sup>38</sup>		828-51-3	1.00	—	-24.9	-23.0	1.9
5	2,7,7-Trimethylbicyclo[3.1.1]heptane-3-carboxylic acid		58096-29-0	0.74	-23.3	-23.5	-21.5	2.1
6	3-Noradamantane carboxylic acid		16200-53-6	0.73	-22.8	-21.9	-17.0	5.0
7	4-(Prop-1-en-2-yl)cyclohexene-1-carboxylic acid		23635-14-5	0.73	-21.5	-21.4	-16.5	4.9
8	2,3,4-Trimethylcyclopentane-1-carboxylic acid		Unknown	0.73	-21.0	No complexation	No complexation	No complexation

**Table 4** Selected guest molecules derived from the virtual screening for query 2

ID	Name	Structure	CAS-No.	Similarity	$\Delta G^\circ$ predicted/ kJ mol <sup>-1</sup>	$\Delta G^\circ$ experimental/ kJ mol <sup>-1</sup>	$\Delta H^\circ$ experimental/ kJ mol <sup>-1</sup>	$T\Delta S^\circ$ experimental/ kJ mol <sup>-1</sup>
2	2-(3-Bicyclo[2.2.1]heptyl)acetic acid <sup>39</sup>		1007-01-8	1.00	—	-20.8	-10.7	10.2
9	(E)-N-(1,7,7-Trimethyl-6-bicyclo[2.2.1]heptylidene)hydroxylamine		2792-42-9	0.79	-21.3	-23.0	-25.2	-2.1
10	5-(Dithiolan-3-yl)pentanoic acid		1077-28-7	0.79	-20.4	-19.6	-15.0	4.7
11	3,7-Dimethyloct-6-enoic acid		18951-85-4	0.75	-21.3	-19.7	-16.5	3.2
12	2-(1-Adamantyl)acetic acid		4942-47-6	0.74	-24.6	-28.8	-24.6	4.2

validation. The correlation  $r^2$  of 0.35 was, in fact, rather low, but increased to 0.65 if the data point of compound **18**, an obvious outlier, is omitted (see Fig. 6). For compound **18** the measured  $\Delta G^\circ$  was clearly weaker than the predicted value.

This discrepancy is attributed to repulsive forces caused by steric interactions due to the branched structure of this guest. The difference between the cross-validated  $r^2$  and the  $r^2$  for the predicted ligands lies with the fact that we only suggested

**Table 5** Selected guest molecules derived from the virtual screening for query 3

ID	Name	Structure	CAS-No.	Similarity	$\Delta G^\circ$ predicted/ kJ mol <sup>-1</sup>	$\Delta G^\circ$ experimental/ kJ mol <sup>-1</sup>	$\Delta H^\circ$ experimental/ kJ mol <sup>-1</sup>	$T\Delta S^\circ$ experimental/ kJ mol <sup>-1</sup>
3	2-[(3-Fluoro-4-phenyl)phenyl]propanoic acid <sup>40</sup>		5104-49-4	1.00	—	-18.8	-23.3	-4.5
13	2-(3-Benzoylphenyl)propanoic acid		22071-15-4	0.94	-17.6	-15.7	-17.1	-1.3
14	4-Phenoxybenzoic acid		2215-77-2	0.91	-17.7	-21.8	-15.8	6.0
15	2-(6-Methoxynaphthalen-2-yl)propanoic acid		22204-53-1	0.91	-17.4	-16.6	-12.6	4.0

**Table 6** Selected guest molecules derived from the virtual screening for query 4

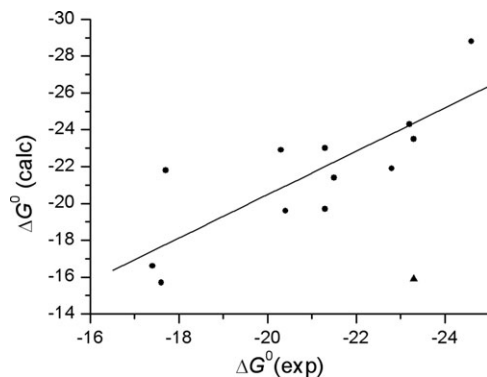
ID	Name	Structure	CAS-No.	Similarity	$\Delta G^\circ$ predicted/ kJ mol <sup>-1</sup>	$\Delta G^\circ$ experimental/ kJ mol <sup>-1</sup>	$\Delta H^\circ$ experimental/ kJ mol <sup>-1</sup>	$T\Delta S^\circ$ experimental/ kJ mol <sup>-1</sup>
4	2-[4-(2-Methylpropyl)phenyl]propanoic acid		15687-27-1	1.00	—	-22.6	-13.5	9.1
16	4- <i>tert</i> -Butylbenzoic acid		98-73-7	0.88	-23.2	-24.3	-20.5	3.8
17	3-(1,2,3,4-Tetrahydronaphthalen-2-yl)propanoic acid		98017-39-1	0.88	-20.3	-22.9	-29.3	-6.4
18	2-Benzyl-3,3-dimethylbutanoic acid		53483-12-8	0.88	-23.2	-15.9	-8.7	7.2

compounds with a high binding affinity for experimental testing. Thus, the overall variance of the data is lower, leading to lower  $r^2$  at equal accuracy. Compound **8** did not show any binding affinity at all. Most probably this is due to the shape of the molecule, which exceeds the diameter of the  $\beta$ -CD cavity (0.6 nm). The size recognition of cyclodextrins is a known phenomenon.<sup>36</sup>

We consider the combination of the similarity-based virtual screening technique and the QSPR model as an efficient way to minimise the drawbacks of each of the two methods when used independently. First, sole application of the similarity tool lacks a concrete estimation of the binding free energy. Although the general principle of similarity is reasonable in many cases, a slight change in structure can have a significant influence on binding properties. For example, binding data of a structural series of similar molecules already investigated in our lab are displayed in Table 7.<sup>37</sup> Despite similarities to 4-*tert*-butylbenzoic acid ranging between 0.96 and 0.83, the corresponding binding free energies differ significantly (between -6.2 and -24.3 kJ mol<sup>-1</sup>). On the other hand, the QSPR model predicts the binding  $\Delta G^\circ$  values with high

correlation ( $r^2 = 0.92$ , RMSE = 3.35 kJ mol<sup>-1</sup>). The application of the QSPR model thus helps to filter out molecules with unfavourable binding energies. In fact, about half of the top-ranking molecules of each of the screening runs were filtered out by the application of the QSPR model.

Conversely, virtual screening based only on the output of a regression model is problematic, because the predictions of QSPR models, such as the one used, are generally only relevant for a limited neighbourhood in chemical space centred around the training set of the model. Thus, the application of the QSPR model alone also leads to an unacceptable number of false positive molecules that do not bind to  $\beta$ -CDs. This results from the fact that the QSPR model was trained only on molecules that bind to  $\beta$ -CD, while non-binding molecules have not been considered. In general, the chemical space of non-binders is too large, that even including negative data into the regression model does not guarantee sufficient consideration of repulsive interactions. Instead, our strategy in this work has been the prior application of the similarity-based screening technique to focus on molecules that both exhibit the principal features of  $\beta$ -CD ligands and lie within the scope of the regression model.

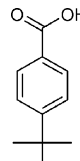
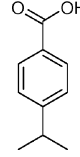
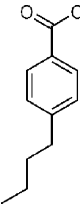
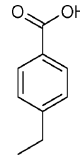
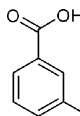
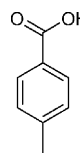
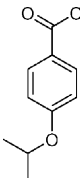


**Fig. 6** Plot of the predicted and experimental  $\Delta G^\circ$  (kJ mol<sup>-1</sup>) values of the screening hits (compounds **5–17** ●, **18** ▲).

## Conclusions

The results validate both the ligand-based screening approach for identifying novel compounds for a given synthetic receptor and the QSPR model for the prediction of binding affinities. While other combinations of similarity screening and regression methods may work as well or even better, the combination used here has been shown to be a promising high throughput alternative to structure-based virtual screenings for the identification of high affinity guests for given receptors. The methodology is much faster compared to docking tools, allowing the screening of very large chemical libraries in a short time on a single CPU without requiring any knowledge

**Table 7** Structural series of benzoic acid derivatives. The similarity was calculated against compound **16**

ID	Name	Structure	CAS-No.	Similarity	$\Delta G^\circ$ predicted/ kJ mol <sup>-1</sup>	$\Delta G^\circ$ experimental/ kJ mol <sup>-1</sup>	$\Delta H^\circ$ experimental/ kJ mol <sup>-1</sup>	$T\Delta S^\circ$ experimental/ kJ mol <sup>-1</sup>
16	4- <i>tert</i> -Butylbenzoic acid		98-73-7	1.00	-23.2	-24.3	-20.5	3.8
19	4-Propan-2-ylbenzoic acid		536-66-3	0.96	-20.6	-19.7	-13.4	6.3
20	4-Butylbenzoic acid		20651-71-2	0.92	-20.1	-21.4	-14.8	6.6
21	4-Ethylbenzoic acid		619-64-7	0.92	-17.2	-15.0	-9.2	5.7
22	3-Methylbenzoic acid		99-04-7	0.88	-13.5	-6.2	-21.0	-14.7
23	4-Methylbenzoic acid		99-94-5	0.88	-15.1	-11.0	-8.0	3.0
24	4-Propan-2-yloxybenzoic acid		13205-46-4	0.83	-16.3	-16.1	-10.6	5.4

of the receptor structure. While  $\beta$ -CD was chosen as a test case because of its technical relevance and the availability of many binding data, the applied methodology can, in principle, be transferred to other systems. The quality of the results will generally depend on the existence of sufficient experimental data for the generation of a reasonably accurate regression model.

### Acknowledgements

We are grateful to Deutsche Forschungsgemeinschaft for funding part of this work (grants AP-101/1 and KA 1804/1).

The authors thank Anne Engelke and Joachim Büch for technical support.

### References

- 1 J. J. Lavigne and E. V. Anslyn, *Angew. Chem., Int. Ed.*, 2001, **40**, 3118.
- 2 M. R. de Jong, R. M. Knegt, P. D. Grootenhuys, J. Huskens and D. N. Reinhoudt, *Angew. Chem., Int. Ed.*, 2002, **41**, 1004.
- 3 T. J. A. Ewing and I. D. Kuntz, *J. Comput. Chem.*, 1997, **18**, 1175.
- 4 F. Corbellini, R. M. A. Knegt, P. D. J. Grootenhuys, M. Crego-Calama and D. N. Reinhoudt, *Chem.-Eur. J.*, 2004, **11**, 298.
- 5 Y. C. Martin, J. L. Kofron and L. M. Traphagen, *J. Med. Chem.*, 2002, **45**, 4350.



- 6 M. Rarey, S. Wefing and T. Lengauer, *J. Comput. Aided Mol. Des.*, 1996, **10**, 41.
- 7 J. A. Grant, M. A. Gallardo and B. T. Pickup, *J. Comput. Chem.*, 1996, **17**, 1653.
- 8 Daylight Chemical Information Systems, (<http://www.daylight.com>).
- 9 MDL keys available from Elsevier MDL, (<http://www.mdli.com>).
- 10 BCI structure fingerprints available from Digital Chemistry, (<http://www.bci.gb.com>).
- 11 A. Kämper, D. Rognan and T. Lengauer, in *Bioinformatics—From Genomes to Therapies*, ed. T. Lengauer, Wiley-VCH, Weinheim, 2007, vol. 2.
- 12 T. Suzuki, M. Ishida and W. M. F. Fabian, *J. Comput. Aided Mol. Des.*, 2000, **14**, 669.
- 13 A. R. Katritzky, D. C. Fara, H. F. Yang, M. Karelson, T. Suzuki, V. P. Solov'ev and A. Varnek, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 529.
- 14 G. Wenz, *Angew. Chem., Int. Ed. Engl.*, 1994, **33**, 803.
- 15 M. V. Rekharsky and Y. Inoue, *Chem. Rev.*, 1998, **98**, 1875.
- 16 K. A. Connors, *Chem. Rev.*, 1997, **97**, 1325.
- 17 M. E. Davis and M. E. Brewster, *Nat. Rev. Drug Discovery*, 2004, **3**, 1023.
- 18 K. Uekama, *Chem. Pharm. Bull.*, 2004, **52**, 900.
- 19 T. Loftsson and D. Duchene, *Int. J. Pharm.*, 2007, **329**, 1.
- 20 H. Tsutsumi, H. Ikeda, H. Mihara and A. Ueno, *Bioorg. Med. Chem. Lett.*, 2004, **14**, 723.
- 21 A. Iaconinoto, M. Chicca, S. Pinamonti, A. Casolari, A. Bianchi and S. Scalia, *Pharmazie*, 2004, **59**, 30–33.
- 22 H. J. Buschmann and E. Schollmeyer, *J. Cosmet. Sci.*, 2002, **53**, 185.
- 23 I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk and V. V. Prokopenko, *J. Comput. Aided Mol. Des.*, 2005, **19**, 453.
- 24 T. Suzuki, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1266.
- 25 MDL<sup>®</sup> ISIS/Draw, [http://www.mdli.com/products/framework/isis\\_draw/index.jsp](http://www.mdli.com/products/framework/isis_draw/index.jsp).
- 26 J. Sadowski and J. Gasteiger, *Chem. Rev.*, 1993, **93**, 2567.
- 27 R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
- 28 A. Steffen, T. Lengauer and J. Apostolakis, QSPR Study on the Predictability of Thermodynamic Properties of Beta-Cyclodextrin Inclusion Complexes, submitted.
- 29 H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola and V. Vapnik, *Advances in Neural Information Processing Systems 9*, ed. M. Mozer, M. I. Jordan and T. Petsche, MIT Press, Cambridge, MA, USA, 1996, p. 155.
- 30 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273.
- 31 C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- 32 M. Ruschhaupt, W. Huber, A. Poustka and U. Mansmann, *Stat. Appl. Genet. Mol. Biol.*, 2004, **3**(1), 37.
- 33 J. Irwin and B. Shoichet, *Abstr. Pap. Am. Chem. Soc.*, 2005, **230**, U1009.
- 34 J. Marialke, R. Körner, S. Tietze and J. Apostolakis, *J. Chem. Inf. Model.*, 2007, **47**, 219.
- 35 E. J. Barker, D. Buttar, D. A. Cosgrove, E. J. Gardiner, P. Kitts, P. Willett and V. J. Gillet, *J. Chem. Inf. Model.*, 2006, **46**, 503.
- 36 A. Müller and G. Wenz, *Chem.-Eur. J.*, 2007, **13**, 2218.
- 37 T. Höfler and G. Wenz, *J. Inclusion Phenom. Mol. Recognit. Chem.*, 1996, **25**, 81.
- 38 J. C. Harrison and M. R. Eftink, *Biopolymers*, 1982, **21**, 1153–1166.
- 39 L. A. Godinez, S. Patel, C. M. Criss and A. E. Kaifer, *J. Phys. Chem.*, 1995, **99**, 17449.
- 40 H. Ueda and J. H. Perrin, *J. Pharm. Biomed. Anal.*, 1986, **4**, 107.