TMACC: Interpretable Correlation Descriptors for Quantitative Structure−Activity Relationships

James L. Melville and Jonathan D. Hirst*

School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom

Received September 25, 2006

Highly predictive topological maximum cross correlation (TMACC) descriptors for the derivation of quantitative structure−activity relationships (QSARs) are presented, based on the widely used autocorrelation method. They require neither the calculation of three-dimensional conformations nor an alignment of structures. We have validated the TMACC descriptors across eight literature data sets, ranging in size from 66 to 361 molecules. In combination with partial least-squares regression, they perform competitively with a current state-of-the-art 2D QSAR methodology, hologram QSAR (HQSAR), yielding larger leave-one-out cross-validated coefficient of determination values (LOO q^2) for five data sets. Like HQSAR, these descriptors are also interpretable but do not requiring hashing. The interpretation both enables the automated extraction of SARs and can give a description in qualitative agreement with more time-consuming 3D and 4D QSAR methods. Open source software for generating the TMACC descriptors is freely available from our Web site.

INTRODUCTION

Quantitative structure−activity relationships (QSARs) are a long-established technique in computer-aided molecular design. As computers have increased in power, new QSAR methodologies have emerged that take into account the three-dimensional structure of molecules, for example, comparative molecular field analysis (CoMFA),[1] or ensembles of three-dimensional structures, for example, the 4D QSAR method of Hopfinger et al.[2] In our own field of interest, catalyst design, we have made use of both 3D and 4D methods.[3,4] However, these techniques are time-consuming, requiring the generation of one or more 3D structures for each molecule and the selection of an optimal alignment for all molecules in the data set. Therefore, faster 2D methods, based only on properties derivable from atom types and connection tables, are still useful and, in a related context, have been shown to outperform 3D techniques.[5] Another desideratum for a QSAR is that the resulting model is interpretable; many QSAR methods have been reported in the literature, but only a few have become popular and been used outside of the groups that originated them.

While the ready availability of software as an easy-to-use application is an important factor, popular QSAR methods also allow visualization, for example, CoMFA, comparative molecular similarity indices analysis (CoMSIA),[6] and grid-independent descriptors (GRIND).[7] The GRIND method is particularly interesting, as it is alignment-independent, obviating the need to superpose molecules in the data set. It works as a 3D analogue of the autocorrelation descriptor,[8] which is a weighted histogram of the atom pairs in a molecular structure. This has the useful property of mapping topological data of a molecule into a fixed length vector, amenable to standard machine learning and statistical methods. Autocorrelation has been extensively used by Gasteiger and co-workers, who have combined it with self-organizing neural networks for QSAR,[9] diversity analysis of combinatorial libraries,[10] and visualization of high-dimensional data.[11] GRIND represents the molecule as a series of distances between pairs of nonbonded interaction terms evaluated on a grid surrounding each molecule, using the GRID molecular mechanics force field.[12] Its major innovation is that it maintains interpretability by only storing one interaction pair at each distance range, that which has the largest absolute value of the product of the two force field interactions. Several studies have shown that the GRIND descriptors are effective in 3D QSAR.[13−15]

Autocorrelation descriptors are still popular in 2D QSAR (see, for instance, refs 16−19 for recent applications), but no GRIND-style interpretation has been attempted with them. Therefore, we have investigated whether the GRIND approach can be used in 2D QSAR. The result of this approach is a set of descriptors, which we call *T*opological *MA*ximum *C*ross-*C*orrelation (TMACC) descriptors. Compared to GRIND, we replace force field interactions measured on a grid with atomic physicochemical values, and the Euclidean 3D distance is replaced with a topological distance based on the shortest bond distance between atoms. Recently, Sutherland et al. (SOW from hereon) published an in-depth study of the behavior of several 2D and 3D QSAR methods against eight data sets, one of the largest and most comprehensive studies to date.[20] They found that the 3D QSAR methods CoMFA and CoMSIA yielded the most predictive models, and of the 2D QSAR methods, only hologram QSAR,[21] (HQSAR) was competitive. Additionally, a recent study by Gedeck and co-workers[22] concluded that 2D fragment-based descriptors such as HQSAR were the best performing descriptors across nearly 1000 corporate data sets. Therefore, to test the TMACC descriptors rigorously, we

* Corresponding author phone: +44-115-951-3478; fax: +44-115-951-3562; e-mail: jonathan.hirst@nottingham.ac.uk.

10.1021/ci6004178 CCC: $37.00
Published on Web 01/20/2007

Table 1. Data Sets Used in this Study, Including Number of Molecules and References for Further Information

data set code	description	number of molecules	ref
ACE	angiotensin converting enzyme (ACE) inhibitors	114	23
ACHE	acetylcholinesterase (AchE) inhibitors	111	20
BZR	benzodiazepine receptor ligands	147	20
COX2	cyclooxygenase−2 inhibitors	282	20
DHFR	dihydrofolate reductase inhibitors	361	20
GPB	glycogen phosphorylase b inhibitors	66	24
THERM	thermolysin inhibitors	76	24
THR	thrombin inhibitors	88	25

validate them against HQSAR using the same eight data sets as those used in the SOW study.

METHODS

Data Sets. To validate the TMACC descriptors, we used eight data sets previously used by SOW to investigate a wide variety of QSAR methods, including HQSAR. These data sets are summarized in Table 1. We combined the training and test sets for this study and discarded inactive molecules (these are only present in data sets BZR, COX2, and DHFR). For all descriptor calculations in this study, we used the molecules in the same protonation state as given in the Supporting Information of ref 20, and we refer readers to that reference for more information.

TMACC Descriptors. The first stage of calculating the TMACC descriptor is to assign a set of numerical values to each atom, representing physicochemical properties of interest. In this study, we used four atomic properties: Gasteiger partial charges,[26] to represent electrostatics; Crippen−Wildman molar refractivity parameters,[27] to represent steric properties and polarizability; Crippen−Wildman log*P* parameters,[27] representing hydrophobicity; and the recently introduced log*S* parameters introduced by Xu and coworkers,[28] representing solubility and, hence, solvation phenomena. For calculating total log*S* values, Xu and coworkers introduced two corrections, for molecular weight and hydrophobicity, but in this study, we were interested in the atomic parameters only and did not apply any corrections to the atomic values. While implementing the log*S* descriptors, we made some minor modifications to the SMARTS strings given in ref 28, in order to get the correct frequencies of atom type for the training set. Further details are given in the Supporting Information. To take account of the different scales used by each set of atomic parameters, we scaled each contribution by the largest absolute value, so that the positive and negative values took maximum values of +1 and −1, respectively. For the Gasteiger partial charges, we took maximum values for positive and negative charges from the "fragmentlike" subset of the ZINC database,[29] consisting of 49 134 molecules, carrying out the calculation with Open Babel 2.0.0.[30,31]

For all data sets, we treated nonpolar hydrogen atoms implicitly, and their atomic value (if any) was added to the value of the heavy atom to which it was bonded. Polar hydrogen atoms were treated explicitly, like any other atom. The standard equation for calculating an autocorrelation descriptor, x_{ac}, is

$$x_{ac}(p,d) = \sum p_i p_j \quad (1)$$

where p is a property (e.g., partial charge) and d is a topological distance between atoms i and j, normally the shortest number of bonds between atoms. The sum is over all atom pairs that are separated by the distance d.

The TMACC descriptor extends this equation in three ways. First, we treat each atomic property that can take positive and negative values as separate properties; for example, we separate partial charge into a positive and negative charge property, similar to the 4D QSAR method of Hopfinger and co-workers.[2] This separates correlations between, for example, two positive charge centers and two negative charge centers into different descriptors. All properties apart from molar refractivity (where all atomic values are positive) were treated in this way. Second, we calculate cross-correlation, as well as autocorrelation values; that is, we allow the property type for atom i to differ from that calculated for atom j. For instance, as well as considering positive charge−positive charge interactions, we also consider positive charge−negative charge interaction and positive charge−negative log*S* interactions. Third, like the GRIND descriptor, we keep only the maximum value calculated for any given distance. The TMACC equation is therefore represented as

$$x_{tmacc}(p,q,d) = \max(p_i q_j, q_i p_j) \quad (2)$$

where q is the property of the second atom. There are two terms to consider for each atom pair, because when $p \neq q$, $q_j p_i \neq p_j q_i$. This procedure produces a real-valued string with a length between 364 and 700 in the data sets used in this study, dependent on the size of the molecules in the data set and the number of properties considered.

For the purposes of interpretation (vide infra), for each descriptor, we record which atoms contribute to the maximum product. In the event of more than one pair having the same value, we record all pairs. The maximum distance for a given data set was as large as the largest distance in any molecule. The minimum distance was zero, that is, we allowed $i = j$. To compare the effect of using only the maximum interaction, we also generated descriptors where, as in standard autocorrelation, we summed the contributions of all pairs. We label these descriptors *T*opological *C*ross-*C*orrelation (TCC) descriptors. All structure handling, atom typing, and descriptor calculation was carried out using the open source Java library JOELib.[32]

The SOW study showed that there was not one best collection of settings for HQSAR. This may also be the case for the TMACC descriptors. To find the best model for the TCC and TMACC descriptors, we used a blockwise backward-stepping technique. We began with a model including all four physicochemical types and then eliminated each physicochemical type sequentially, removing all cross-term blocks also. If any reduced model resulted in the q^2 increasing, we treated this as the new model. There is a slight risk of overfitting with this method, but we never attempted to remove more than one property from each model, leading to only five comparisons in each data set, which is orders of

magnitude smaller than the number of comparisons made using typical variable selection techniques such as simulated annealing and genetic algorithms. This small number of comparisons is similar to the number of comparisons used to choose the best HQSAR model (four in this case) and should therefore not give any unfair advantage to the TMACC method.

Hologram Descriptors. The SOW study indicated that HQSAR was the most effective of the 2D QSAR techniques that were investigated. Therefore, we decided to compare the TMACC descriptor with HQSAR. We wrote our own implementation to produce holograms, rather than those available in the Sybyl program (Tripos Inc., St Louis, Missouri), so we would not expect to see identical results. However, the technique should be sufficiently general that successful application is not tied to a particular implementation. All molecules were treated as hydrogen-suppressed graphs. Next, all connected subgraphs of size 4−7 (the sizes used in the SOW study) were generated, using an algorithm similar to that described by Rücker and Rücker.[33] Each subgraph was represented as a SMARTS string.[34] The specificity of the SMARTS properties was chosen to match those used by SOW: we included the atomic number (A); the bond type connecting each atom in the subgraph (B); the total connectivity of each atom in the subgraph, that is, the number of neighbors of the atom in the molecule, not the subgraph (Co); the chirality, using the lexical, string-based definition used in isomeric SMILES,[35] not the absolute R/S configuration (C); and the number of hydrogens attached to each atom (H). Each SMARTS fragment was then canonicalized. This was achieved by labeling each atom with the properties above as a string and using Java's string comparison methods to assign priority. If atoms were tied, the (sorted) properties of the neighbor atoms were concatenated to the string and the comparison repeated. If there were still ties, the next nearest neighbor properties were concatenated to the string and the process repeated until the ties were broken, or there were no further neighbors to add, in which case the atoms are identical according to the level of representation chosen. A count of each canonicalized SMARTS string was recorded for every molecule, and these were used as descriptors. Because of the large number of structures that are generated in this process, it is usual to "fold" the fingerprint, by hashing it to a much smaller length. However, this loss of information can have a deleterious effect, and Seel et al. recommended that unhashed strings be used.[36] In the SOW study, they used a hash length of between 1500 and 4500 to minimize this effect. In our study, the length of the hologram varied between 3000 and 18 000, and we used no hashing at all. We used JOELib to develop the hologram-generation program in Java.

Validation. Descriptors with zero variance were removed, followed by QSAR model building with partial least-squares (PLS) regression,[37] using the SIMPLS algorithm of de Jong.[38] The program was written in Java, using a modification of the open source machine-learning library Weka.[39] For the TMACC descriptors, those with the same physicochemical type correlation, $p_i q_j$, but different bond distances, d_{ij}, were treated as a single block, and block-scaling was applied, to take into account the differences in scale between the different atom types, similar to how different fields are scaled in 3D QSAR techniques such as CoMFA. Autoscaling was used for the HQSAR descriptors. The optimum number of components (ONC) for each data set was established by cross-validation. It is common to use an external test set to establish an estimate of predictive accuracy.[40] However, results presented by Hawkins and co-workers[41,42] and by Faber[43] suggest that the external test set must be large to give results as reliable as those given by cross-validation. Variable selection complicates the issue, making cross-validation less reliable. However, as our variable selection is very mild, this should not bias the results significantly. We repeated some of the work presented below using the training/test set partition provided by SOW and achieved qualitatively similar results, which are reported in the Supporting Information. Therefore, we use cross-validation only in this study, as this makes more use of the data for model building and should therefore provide more reliable results.

For all data sets, we used leave-one-out (LOO) cross-validation, which the SOW study and those of Hawkins and co-workers have identified as being optimal for determining predictive accuracy. The cross-validated coefficient of determination, q^2, is defined as

$$q^2 = 1 - \frac{\mathrm{PRESS}}{\mathrm{TSS}} \tag{3}$$

where the predicted residual sum of squares, PRESS, is

$$\mathrm{PRESS} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \tag{4}$$

and the total sum of squares, TSS, is given by

$$\mathrm{TSS} = \sum_{n=1}^{N} (y_n - \bar{y}_{/g})^2 \tag{5}$$

y_n is the experimental activity of molecule n, $\hat{y}_n$ is the predicted activity, and N is the total number of molecules in the data set. For the calculation of the TSS, it is normal to use $\bar{y}$, the mean experimental activity of all molecules in the data set. We have used a slight modification of this formula, making use of $\bar{y}_{/g}$, where $/g$ represents the molecules not in the cross-validation group g being predicted (for LOO, g consists only of the one left-out observation). This has only a minor effect on the q^2 values but leads to PRESS = TSS for the null model, where the activity of each molecule is assigned to be the mean of the activity of the molecules used to build the model during cross-validation. This naturally leads to a definition of $q^2 = 0.0$ for the null model. This definition of TSS is identical to that used by other workers, when generating the final, non-cross-validated model. The optimum model was that which yielded the first minimum in s_{PRESS},

$$s_{\mathrm{PRESS}} = \sqrt{\frac{\mathrm{PRESS}}{N - A - 1}} \tag{6}$$

where A is the number of components. No rescaling was carried out during cross-validation. The coefficient of determination for the final, non-cross-validated model, R^2, is defined analogously to q^2, except that the estimated activities, $\hat{y}_n$, are the non-cross-validated values.

Scramble Set. Scramble sets are frequently used to validate QSAR methods and provide a measure of the risk of chance correlations occurring. Therefore, for each data set, we carried out the following procedure. First, the activities were shuffled randomly. A model was then built in the usual way, followed with variable selection by dropping properties. The entire process was repeated 1000 times. We report the average q^2, R^2, and ONC over the 1000 runs.

Baseline Comparison. When assessing the adequacy of a QSAR, some authors advocate a cutoff based on LOO q^2; for example, Tropsha and co-workers suggest 0.50.[40] However, the q^2 is very much data-set-dependent, in particular, on the range of experimental activities. For example, for the GPB data set considered here, obtaining a q^2 of 0.65 would seem like a good result; this corresponds to a root-mean-square (RMS) error in prediction of 0.67 pK_i units. However, the same q^2 value for the THERM data set would only yield an RMS error of 1.21 pK_i units, which may well be considered insufficiently accurate. Therefore, care must be taken when interpreting q^2 values. In the fields of docking and similarity searching, there have been attempts to prevent overly optimistic interpretation of results by ensuring that inactive molecules had a similar property profile to that of actives[44] or by comparing very simple descriptors with more sophisticated fingerprint techniques.[45] For this study, we have adopted the practice of Bender and Glen,[45] by using a set of "dumb" atom count descriptors, where each molecule is represented by a fingerprint of length 12, providing a count of the number of boron, carbon, nitrogen, oxygen, fluorine, phosphorus, sulfur, chlorine, bromine, iodine, non-hydrogen, and all atoms in the molecule.

HQSAR Validation. As we have not used the Tripos implementation of HQSAR, we checked that our implementation gave results similar to those used in the SOW paper. For this part of the study only, we used the training set molecules specified in ref 20. q^2 was calculated with the more usual definition of the total sum of squares:

$$\mathrm{TSS} = \sum_{n=1}^{N} (y_n - \bar{y})^2 \qquad (7)$$

where $\bar{y}$ is the mean experimental activity of *all* molecules in the data set.

Interpretation. To interpret the TMACC descriptors, we rescale the coefficients from the non-cross-validated model to reverse the effects of block scaling. The regression constant is ignored. For each unscaled descriptor in the model, x_i, we calculate the partial activity contributed by $\beta_i x_i$, where β_i is the unscaled regression coefficient. Each atom that contributes to the descriptor then gets an equal share of the partial activity. By this scheme, the activity is partitioned among the atoms in the molecule. It is then possible to visualize which atoms contribute most to the activity of each molecule, by coloring each atom according the sign and magnitude of its partial activity. Diagrams were produced with ISIS/Draw (Elsevier MDL, San Ramon, CA).

RESULTS AND DISCUSSION

Baseline Comparison. To establish a baseline for the predictivity of our descriptors, we first derived models using the "dumb" simple atom-count descriptors. Additionally, we also built PLS models using the 2.5D descriptors provided in the Supporting Information of ref 20. Results for the "dumb" and 2.5D descriptors are given in the first two columns of Table 2. For two data sets (GPB and THR), the use of the "dumb" descriptors results in an extremely unpredictive model. For most of the others, the resulting models have $q^2 \leq 0.30$, which would be considered unlikely to be of use. However, use of the "dumb" descriptors results for the ACE data set gives a q^2 of 0.55, much larger than the other results. This may suggest that it is easier to produce large q^2 values for the ACE data set than it is for the others.

Table 2. q^2 Values for Different Descriptors against the Eight Data Sets Used in this Study

data set	dumb	2.5D	TMACC	TCC	HQSAR
ACE	0.55	0.69	0.71	0.67	0.63
ACHE	0.23	0.27	0.62	0.47	0.58
BZR	0.17	0.34	0.41	0.40	0.40
COX2	0.14	0.42	0.54	0.54	0.49
DHFR	0.20	0.58	0.60	0.61	0.74
GPB	0.07	0.46	0.62	0.56	0.58
THERM	0.30	0.36	0.56	0.46	0.71
THR	0.09	0.38	0.64	0.65	0.66

Table 3. Comparison between TCC and TMACC Descriptors.

	TCC		TMACC	
data set	q^2	atom types[a]	q^2	atom types[a]
ACE	0.67	ERS	0.71	EHRS
ACHE	0.47	HRS	0.62	EHRS
BZR	0.40	HRS	0.41	EHRS
COX2	0.54	EHRS	0.54	EHS
DHFR	0.61	EHRS	0.60	EHRS
GPB	0.56	EHS	0.62	EHRS
THERM	0.46	EHS	0.56	EHS
THR	0.65	EHRS	0.64	HRS

[a] Meaning of atom type codes: E, electrostatic; H, hydrophobicity; R, molar refractivity; S, solubility.

Effect of Retaining Maximum Interactions Only. Having established threshold q^2 values for success using the "dumb" descriptors, we turned our attention to the TMACC descriptors. The first question we sought to answer was whether only keeping the maximum product for each correlation had a deleterious effect on the QSARs obtained. Therefore, for each data set, we generated PLS models using TMACC and TCC descriptors. The best models found are given in Table 3, with the q^2 values reproduced from Table 2, along with one-letter codes to indicate which atom types were retained in the final model. Across the eight data sets, the average q^2 value for the TMACC descriptor is much larger than that of the dumb descriptors and larger than that of the 2.5D descriptors. However, because of differences in size, composition, and difficulty in establishing QSARs for the different data sets, it is more useful to compare the number of times a given descriptor outperforms a reference. On this basis, the table shows that both TCC and TMACC descriptors are superior to the dumb descriptors for all data sets. Compared to the 2.5D descriptors, the TCC descriptors are better for seven of the eight data sets, while the TMACC is again superior across all eight data sets. This establishes that the correlation descriptors can perform well for these data sets. Perhaps surprisingly, Table 3 also indicates that

Table 4. Comparison of Hologram QSAR Results from Ref 20 and the Implementation Used in this Study

	ref 20		this study	
data set	optimum q^2	optimum representation[a]	optimum q^2	optimum representation[a]
ACE	0.72	ABCoHC	0.67	ABCoC
ACHE	0.34	ABCoH	0.49	ABCoHC
BZR	0.42	ABCoC	0.47	ABCoH
COX2	0.50	ABCo	0.50	ABCoC
DHFR	0.69	ABCoHC	0.69	ABCo
GPB	0.66	ABCo	0.62	ABCo
THERM	0.49	ABCo	0.66	ABCoC
THR	0.50	ABCoH	0.68	ABCoC

[a] Representation code: A, atoms; B, bonds; Co, connectivity; H, number of hydrogens; C, chirality.

the TCC descriptor is superior to the TMACC descriptors for only two of the eight data sets. That is, despite only keeping one correlation at each bond distance, there has been essentially *no* loss of information in the descriptors. Indeed, the TMACC descriptor outperforms the TCC descriptor for the majority of data sets. This may seem counterintuitive, given that the TCC descriptor contains information about more atom pairs than the TMACC. We suggest the following explanation. The TMACC descriptors retain only the largest interaction values, which are likely to be those most relevant for molecular recognition; the additional interactions used in TCC only add noise to the descriptor. For five of the data sets, the TMACC models consist of all four of the atom properties generated. This suggests that all atom types are giving useful information. The log*S* descriptor is included in all models, suggesting that it is supplying useful information beyond that supplied by the other descriptors. To quantify this, we concatenated the log*P*+:log*P*−, log*P*+:log*P*+, and log*P*−:log*P*− TMACC descriptors for each molecule in each data set and measured the correlation between the vector and that generated using log*S*. The correlation was moderate, with a maximum Pearson correlation, *r*, of 0.69 for the ACE, THERM, and THR data sets. However, the maximum correlation between an analogously generated partial charge vector and log*P* was 0.64, a comparable value, and it is widely acknowledged that partial charge and atomic log*P* can contain complementary information in QSAR studies, so atomic log*S* values may therefore prove useful in other QSAR applications, for example, CoMFA.

HQSAR Validation. Before comparing the TMACC descriptors with the HQSAR descriptors, we wished to ensure that our implementation could produce models of comparable quality. We therefore followed the same procedure as that carried out in ref 20 and compared the resulting q^2 values. Results are shown in Table 4. Our implementation of HQSAR does not yield the same optimal q^2 or representation, but while our results are not identical to those obtained by SOW, they are similar for most data sets, with a difference in the LOO q^2 of no larger than 0.05 units for five data sets. For the three data sets with a larger difference in q^2 (ACHE, THR, and THERM), the models produced with our implementation are more predictive. Therefore, our HQSAR implementation provides a fair comparison with the TMACC descriptors.

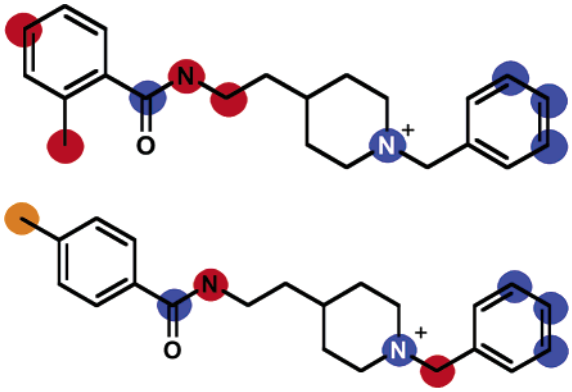

Figure 1. Visualization of the TMACC/PLS model showing the effect of modifying an ortho-methyl-substituted AChE inhibitor (molecule 2-3b, top, $pIC_{50} = 6.00$) to a para-methyl-substituted compound (molecule 2-3d, bottom, $pIC_{50} = 6.74$). Highlighted atoms indicate those that are most important for increasing (blue) and decreasing (red) the potency of the inhibitors. The orange color indicates that the methyl substitution becomes less potency-reducing in the para substitution, in line with a previously published SAR.

Comparison with HQSAR. The SOW study concluded that HQSAR was the only 2D descriptor studied that was competitive with the 3D QSAR methods (CoMFA and CoMSIA). Therefore, we test the ability of the TMACC descriptors against HQSAR as a stringent measure of their usefulness. HQSAR values are given in the final column of Table 2. Comparing the two descriptors by counting the number of "victories" for the TMACC descriptor, the TMACC descriptors are superior for six of the eight data sets. These promising results suggest that the TMACC descriptor is competitive with the best 2D QSAR methods currently available.

Scramble Set. We carried out a scramble set test, by repeating our model build procedure for 1000 scrambled responses for each data set. For all data sets, the average q^2, R^2, and ONC were never larger than 0.01, 0.03, and 0.19, respectively. These values are all very close to zero and substantially smaller than the unscrambled results, which gives us greater confidence that our results are valid and not due to chance correlation.

Interpretation. An important property of any descriptor for QSAR is that it be interpretable. The TMACC descriptors, when combined with a linear regression method such as PLS, can be used to partition the predicted activity of a molecule among its constituent atoms. To demonstrate the ability of the TMACC descriptor to provide useful interpretation, we examine a subset of the SOW benchmark data sets.

The AChE data set represents a collection of piperidine derivatives reported by Sugimoto and co-workers.[46−48] For 1-benzyl-4-[2-(*N*-benzoylamino)ethyl]piperidine derivatives, it was noted that activity increased upon moving from an ortho-substituted phenyl ring to para substitution, for both methyl- and nitro-substituted compounds.[46] Using the labeling of ref 20, Figure 1 displays molecules 2-3b and 2-3d, the ortho- and para-methyl-substituted compounds, respectively. We display the atoms that were identified by the TMACC descriptors to most contribute positively to activity in blue and those that decrease activity the most in red. Additionally, we display the coloring of the methyl group. In the interest of clarity, we do not show the contributions of the other atoms. When the ortho-substituted compound

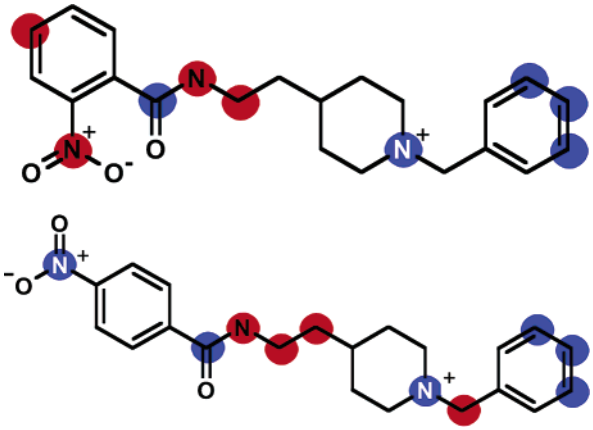

Figure 2. Visualization of the TMACC/PLS model showing the effect of modifying an ortho-nitro-substituted AChE inhibitor (molecule 2-3e, top, $pIC_{50} = 6.06$) to a para-nitro-substituted compound (molecule 2-3g, bottom, $pIC_{50} = 7.26$). Highlighted atoms indicate those that are most important for increasing (blue) and decreasing (red) the potency of the inhibitors. The shift from red to blue coloring of the nitro group again highlights the increase in potency that this change engenders—compare with Figure 1.

is examined, the contribution of the ortho-methyl group is colored red; hence, this group has been identified as an activity-decreasing group. However, upon examination of the para-substituted group, the methyl group is colored orange; this also indicates an activity-decreasing group, but its effect is more moderate. The other atoms of the molecule that have been identified as being particularly influential on activity are fairly constant between the two molecules; hence, we can clearly assign the increase in activity in going from the ortho to the para substitution to the movement of the methyl group. Figure 2 shows a similar change from ortho to para substitution, but this time with a nitro group, illustrated by molecules 2-3e and 2-3g. Again, the ortho substitution is indicated as being activity-reducing, while para substitution improves the activity. However, the change in coloring for this substitution indicates that it has a more dramatic effect on activity: the methyl substitution changed from red to orange, indicating a change from activity-decreasing to moderately activity-decreasing; here, the nitro substitution has changed from red to blue, indicating a change from activity-decreasing to activity-increasing. This is borne out by examining the pIC_{50} values of the compounds; changing from an ortho-methyl to a para-methyl substitution increases the pIC_{50} by 0.74 log units, while changing from ortho-nitro to para-nitro substitution increases the activity by 1.20 log units. It is also pleasing that essentially the same atoms are identified as being important for activity for both the methyl-substituted and nitro-substituted molecules; for example, the amide carbonyl carbon, the piperidine nitrogen, and the same three carbon atoms of the benzene ring are identified as important activity enhancers in all four molecules. These observations can be rationalized by noting that replacement of the amide with an amine results in an almost complete disappearance of activity, while replacement of the N-benzyl group with a hydrogen atom or a cyclopropylmethyl group results in a large decrease in potency, indicating the requirement for a full six-membered ring.[46] Additionally, both the ortho-substituted molecules clearly indicate that a lack of substitution at the 4 position of the benzene ring reduces activity.

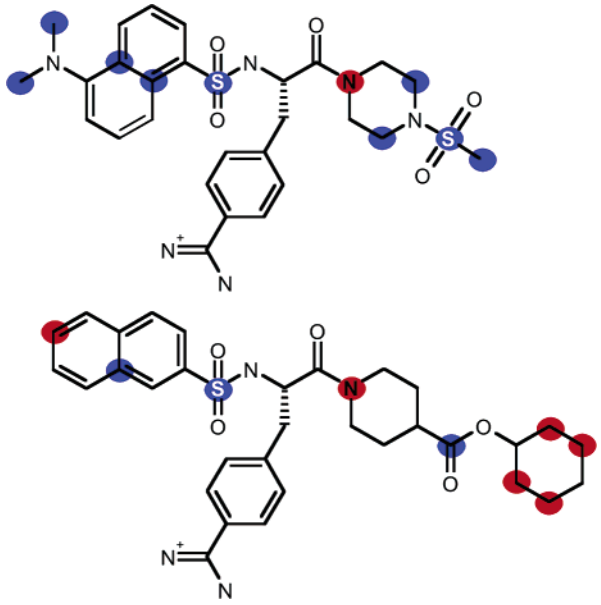

Figure 3. Visualization of the TMACC/PLS QSAR model derived for two molecules from the thrombin data set. The top molecule is a potent inhibitor (molecule 01, $pK_i = 8.38$), the lower molecule a less potent inhibitor (molecule 58, $pK_i = 6.12$). The atoms with the largest influence on the activity are highlighted in blue (most activity-enhancing) and red (most activity-decreasing).

As a second example, we were interested to see if our TMACC interpretation was consistent with 3D QSAR techniques. We chose the thrombin (THR) data set studied by Böhm et al.[25] with the CoMSIA method, for which they published a contour plot superimposed over the bioactive conformations of two thrombin inhibitors, indicating areas of the binding site which should be kept occupied and unoccupied to maximize binding affinity. Figure 3 shows the two inhibitors. The first is an inhibitor with high affinity (molecule 01, top) with a measured pK_i of 8.38, and the second is a less active inhibitor (molecule 58, bottom), which had a measured pK_i of 6.12. Ignoring the 3-amidinopheny-lalanine scaffold that was part of all inhibitors in this data set, we highlight those atoms that our TMACC/PLS analysis indicated contributed the most to increasing activity (blue) and decreasing activity (red). We compared this to the first panel in Figure 12 of ref 25, which shows the result of applying CoMSIA to these molecules.

Böhm et al. noted two regions relevant for these two inhibitors, one that increased activity, the other decreasing activity. The activity-increasing region is occupied by half of the piperazine ring and the sulfonamide group (molecule 01) and half the piperidine ring and the ester group (molecule 58). It can be seen from Figure 3 that the TMACC also highlights the piperazine ring and the sulfonamide group in molecule 01, and the carbonyl carbon of molecule 58 is also highlighted as activity-increasing. The CoMSIA analysis revealed an activity-decreasing region close to the cyclohexyl group of molecule 58, which extends out beyond molecule 01, and the TMACC descriptors also highlight the cyclohexyl ring as being deleterious to activity. The CoMSIA study failed to highlight the importance of hydrophobic interactions due to the aromatic sulfonamide group in the S3 pocket for thrombin, although these were found for the same set of molecules for trypsin and Factor Xa; given that the S3 pocket of thrombin has two hydrophobic residues (Leu99 and

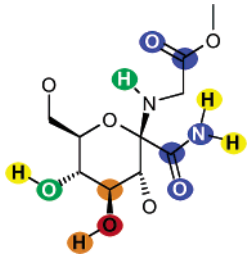

Figure 4. Visualization of the TMACC/PLS model for the inhibition of glycogen phosphorylase (GPB), applied to an active glucose analogue inhibitor ($pK_i = 4.80$). Atoms have been highlighted that were deemed to be involved in important ligand−protein interactions in a previous 4D QSAR study, see Table 2 and Figure 4 of ref 49. Atoms are color-coded by their contribution to the activity: red, strongly activity-decreasing; orange, moderately activity-decreasing; yellow, neutral; green, moderately activity-increasing; blue, strongly activity-increasing.

Ile174) while trypsin has only one (Leu99), it seems likely that hydrophobic interactions are at least as important in thrombin in this part of the receptor. It is therefore probably significant that the TMACC descriptor highlights atoms in the hydrophobic naphthalene ring of both molecules. The N-methyl substitution is also highlighted as important in molecule 01; similarly, the lack of substitution in that region is shown to be activity-reducing for molecule 58. One observation that may appear unusual is that the heterocycle/amide nitrogen is considered to be activity-decreasing, despite the majority of the data set having a nitrogen atom at that position. However, a closer look at the data shows that there are three pairs of molecules in the data set where the amide has been replaced by an ester, without any further structural modification. In all cases, there is a significant increase in activity. This can be seen by comparing molecules 63 and 41 in ref 25, where replacing a methylamino group with a methoxy group increases activity by 1.8 log units; molecules 64 and 28, where an isopropylamino substitution is replaced by an isopropoxy group, increasing activity by 1.3 log units; and molecules 31 and 74, where replacement of a butylamino group with a butoxy group increases activity by 1.0 log unit. Thus, the TMACC descriptor provides structural insight consistent with, and complementary to, the 3D QSAR analysis.

For a final example, we compared our TMACC interpretation for the set of glucose analogue inhibitors of glycogen phosphorylase-b (GPB)[24] to the analysis of Pan et al.,[49] who employed a receptor-dependent 4D QSAR method (RD-4D-QSAR). RD-4D-QSAR requires a molecular dynamics simulation to be carried out for each receptor−ligand complex. As our method is purely two-dimensional in nature without using any protein structural data, comparing the two approaches represents a stern test of the robustness of the TMACC descriptors. Pan et al. identified a "functional" region of the protein where 11 significant protein−ligand interactions (termed grid cell occupancy descriptors, GCODs, in ref 49) were highlighted. As these could be mapped to particular functional groups of the inhibitor, we compared these GCODs to the atoms in molecule 45 (using the numbering in ref 24; this corresponds to molecule 24 in ref 49), shown in Figure 4. This molecule was also used in the RD-4D-QSAR study to illustrate the interactions. We display polar hydrogens in this figure, where the work of Pan et al. indicated that these were involved in receptor−ligand interactions. An increase in inhibition potency was indicated for the proton and the carbonyl group of the amide in the β substituent of the inhibitor. Figure 4 shows that the proton is labeled in green (indicating a moderately positive effect on potency) and the carbonyl group in blue (a strongly positive effect on potency), in agreement with the GCOD analysis. For the alpha substituted amide, the carbonyl group was also shown to be responsible for an activity increase; again, we see that the TMACC descriptors have labeled this as strongly activity-increasing. The protons of the amide group were also suggested to increase potency by hydrogen bonding. In contrast, our 2D model labels the protons as having a neutral effect on potency; however, the nitrogen is shown to be strongly activity-enhancing, so taken together, the NH_2 group can be shown to be in agreement with the 4D QSAR model. On the glucose ring, the C_3−OH group was suggested to decrease activity, because of a conformational change inducing an unfavorable interaction. The TMACC descriptor is again in agreement, labeling the entire C−OH group as activity-decreasing. However, it was suggested by the 4D QSAR model that the C_3 atom itself was involved in a favorable interaction, while the TMACC descriptor was unable to highlight this behavior. Finally, the C_4−OH group is marked as being involved in both favorable and unfavorable interactions in the 4D QSAR model, and the TMACC model also reflects this ambiguity, where the oxygen is marked as being moderately activity-increasing, but the hydroxyl proton does not contribute markedly to the potency of this inhibitor.

To summarize the above discussion, we have shown that the TMACC QSARs lend themselves to simple and intuitive interpretations, as well as having excellent predictive abilities, and can reproduce the features of both manually extracted structure−activity relationships, as well as more time-consuming 3D and 4D QSAR methods.

CONCLUDING REMARKS

The TMACC descriptors represent a simple but effective modification of 2D autocorrelation descriptors, analogous to that used in the 3D GRIND descriptor. We have validated the TMACC with eight data sets and found that allowing cross-correlations between different atom types substantially enhances performance. Perhaps surprisingly, retaining only the maximum correlation value for each atom pair does not have a deleterious effect on the results and, in most cases gives superior results. In addition to the widely used Crippen−Wildman atom types for molar refractivity and hydrophobicity, and Gasteiger partial charges, the recently introduced solubility atom types described by Xu and coworkers provide significant value. Comparison to an implementation of HQSAR demonstrates that the TMACC performs competitively and, in the majority of cases, is marginally superior. Like HQSAR, it is possible to interpret the QSAR visually, providing insight in agreement with 3D and 4D QSAR methods. However, TMACC descriptors are of a much lower dimensionality than HQSAR and do not require hashing. The TMACC descriptor can also be easily extended by the introduction of different atom types, for example, more sophisticated partial charge calculations. Like any QSAR descriptor, the TMACC is not a magic bullet—it does not perform universally well on all the data sets studied

here. One current weakness of the TMACC descriptor is that it is currently insensitive to chirality. Additionally, the use of physicochemical cross-correlations means that identifying which properties are responsible for activity is difficult. By removing the interproperty cross-correlations (e.g., not including log*P*−log*S* cross-correlations), such an interpretation is possible, with a concomitant decrease in descriptor length. However, in our experiments, the quality of the QSARs was also lowered. Nonetheless, we feel that the interpretation provided by the fully cross-correlated TMACC descriptors is valuable and can be seen as complementary to other QSAR methods that focus on physicochemical properties (e.g., CoMFA). Nonetheless, we hope to address some of these shortcomings in future work. Source code (in Java) to generate the TMACC descriptors is freely available from our Web site under the GNU General Public License at http://comp.chem.nottingham.ac.uk/download/tmacc/index.html.

ACKNOWLEDGMENT

We thank the EPSRC for support (GR/S75765/01) and an equipment grant (GR/R62052/01) for computers. We are also grateful for the use of the High Performance Computing Facility at the University of Nottingham.

Supporting Information Available: Details of modifications to the log*S* descriptor and training/test set statistics. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−67.

(2) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(3) Melville, J. L.; Andrews, B. I.; Lygo, B.; Hirst, J. D. Computational Screening of Combinatorial Catalyst Libraries. *Chem. Commun.* **2004**, 1410−1411.

(4) Melville, J. L.; Lovelock, K. R. J.; Wilson, C.; Allbutt, B.; Burke, E. K.; Lygo, B.; Hirst, J. D. Exploring Phase-Transfer Catalysis with Molecular Dynamics and 3D/4D Quantitative Structure−Selectivity Relationships. *J. Chem. Inf. Model.* **2005**, *45*, 971−981.

(5) Brown, R. D.; Martin, Y. C. Use of Structure Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(6) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−46.

(7) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(8) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−60.

(9) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(10) Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing Similarity and Diversity of Combinatorial Libraries by Spatial Autocorrelation Functions and Neural Networks. *Angew. Chem., Int. Ed. Engl.* **1996**, *34*, 2674−2677.

(11) Bienfait, B.; Gasteiger, J. Checking the Projection Display of Multivariate Data with Colored Graphs. *J. Mol. Graphics Modell.* **1997**, *15*, 203−215.

(12) Goodford, P. J. A Computational-Procedure for Determining Energetically Favorable Binding-Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(13) Carosati, E.; Lemoine, H.; Spogli, R.; Grittner, D.; Mannhold, R.; Tabarrini, O.; Sabatini, S.; Cecchetti, V. Binding Studies and GRIND/ALMOND-Based 3D QSAR Analysis of Benzothiazine Type K-ATP-Channel Openers. *Bioorg. Med. Chem. Lett.* **2005**, *13*, 5581−5591.

(14) Cianchetta, G.; Li, Y.; Kang, J. S.; Rampe, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. Predictive Models for hERG Potassium Channel Blockers. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3637−3642.

(15) Sciabola, S.; Carosati, E.; Baroni, M.; Mannhold, R. Comparison of Ligand-Based and Structure-Based 3D-QSAR Approaches: A Case Study on (Aryl-)Bridged 2-Aminobenzonitriles Inhibiting HIV-1 Reverse Transcriptase. *J. Med. Chem.* **2005**, *48*, 3756−3767.

(16) Fernandez, M.; Tundidor-Camba, A.; Caballero, J. M. 2D Autocorrelation Modeling of the Activity of Trihalobenzocycloheptapyridine Analogues as Farnesyl Protein Transferase Inhibitors. *Mol. Simul.* **2005**, *31*, 575−584.

(17) Prabhakar, Y. S.; Rawal, R. K.; Gupta, M. K.; Solomon, V. R.; Katti, S. B. Topological Descriptors in Modeling the HIV Inhibitory Activity of 2-Aryl-3-pyridyl-thiazolidin-4-ones. *Comb. Chem. High Throughput Screen.* **2005**, *8*, 431−437.

(18) Spycher, S.; Pellegrini, E.; Gasteiger, J. Use of Structure Descriptors to Discriminate between Modes of Toxic Action of Phenols. *J. Chem. Inf. Model.* **2005**, *45*, 200−208.

(19) Rhodes, N.; Clark, D. E.; Willett, P. Similarity Searching in Databases of Flexible 3D Structures Using Autocorrelation Vectors Derived from Smoothed Bounded Distance Matrices. *J. Chem. Inf. Model.* **2006**, *46*, 615−619.

(20) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(21) Hurst, J. R.; Heritage, T. W. Molecular Hologram QSAR. U.S. Patent 5,751,605, 1996.

(22) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model*. **2006**, *46*, 1924−1936.

(23) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−84.

(24) Gohlke, H.; Klebe, G. DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J. Med. Chem.* **2002**, *45*, 4153−4170.

(25) Böhm, M.; Stürzebecher, J.; Klebe, G. Three-Dimensional Quantitative Structure−Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis To Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and Factor Xa. *J. Med. Chem.* **1999**, *42*, 458−477.

(26) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity: A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−22.

(27) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(28) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.

(29) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(30) The Open Babel Package, version 2.0.0. http://openbabel.sourceforge.net/ (accessed Dec 1, 2005).

(31) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk−Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991−998.

(32) JOELib. http://sourceforge.net/projects/joelib/ (accessed Dec 1, 2005).

(33) Rücker, G.; Rücker, C. Automatic Enumeration of All Connected Subgraphs. *MATCH* 2000, 145−149.

(34) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Jun 1, 2006).

(35) Daylight Theory: SMILES. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed Jun 1, 2006).

(36) Seel, M.; Turner, D. B.; Willett, P. Effect of Parameter Variations on the Effectiveness of HQSAR Analyses. *Quant. Struct.-Act. Relat.* **1999**, *18*, 245−252.

(37) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(38) de Jong, S. SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251−63.
(39) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.
(40) Golbraikh, A.; Tropsha, A. Beware of *q*2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.
(41) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(42) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.
(43) Faber, N. K. M. Estimating the Uncertainty in Estimates of Root Mean Square Error of Prediction: Application to Determining the Size of an Adequate Test Set in Multivariate Calibration. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 79−89.
(44) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein−Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Model.* **2004**, *44*, 793−806.
(45) Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.
(46) Sugimoto, H.; Tsuchiya, Y.; Sugumi, H.; Higurashi, K.; Karibe, N.; Iimura, Y.; Sasaki, A.; Kawakami, Y.; Nakamura, T.; Araki, S.; Yamanishi, Y.; Yamatsu, K. Novel Piperidine-Derivatives − Synthesis and Antiacetylcholinesterase Activity of 1-Benzyl-4-[2-(N-benzoylamino)ethyl]piperidine Derivatives. *J. Med. Chem.* **1990**, *33*, 1880−1887.
(47) Sugimoto, H.; Tsuchiya, Y.; Sugumi, H.; Higurashi, K.; Karibe, N.; Iimura, Y.; Sasaki, A.; Araki, S.; Yamanishi, Y.; Yamatsu, K. Synthesis and Structure−Activity Relationships of Acetylcholinesterase Inhibitors − 1-Benzyl-4-(2-phthalimidoethyl)piperidine and Related Derivatives. *J. Med. Chem.* **1992**, *35*, 4542−4548.
(48) Sugimoto, H.; Iimura, Y.; Yamanishi, Y.; Yamatsu, K. Synthesis and Structure−Activity Relationships of Acetylcholinesterase Inhibitors − 1-Benzyl-4-[(5,6-dimethoxy-1-oxoindan-2-yl)methyl]piperidine Hydrochloride and Related Compounds. *J. Med. Chem.* **1995**, *38*, 4821−4829.
(49) Pan, D. H.; Liu, J. Z.; Senese, C.; Hopfinger, A. J.; Tseng, Y. Characterization of a Ligand-Receptor Binding Event Using Receptor-Dependent Four-Dimensional Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem.* **2004**, *47*, 3075−3088.

CI6004178