Exploring Phase-Transfer Catalysis with Molecular Dynamics and 3D/4D Quantitative Structure−Selectivity Relationships

James L. Melville,[†] Kevin R. J. Lovelock,[†] Claire Wilson,[†] Bryan Allbutt,[†] Edmund K. Burke,[‡] Barry Lygo,[†] and Jonathan D. Hirst*[,†]

School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, U.K., and School of Computer Science & Information Technology, University of Nottingham, Jubilee Campus, Nottingham NG8 2BB, U.K.

Received February 11, 2005

Quantitative Structure−Selectivity Relationships (QSSR) are developed for a library of 40 phase-transfer asymmetric catalysts, based around quaternary ammonium salts, using Comparative Molecular Field Analysis (CoMFA) and closely related variants. Due to the flexibility of these catalysts, we use molecular dynamics (MD) with an implicit Generalized Born solvent model to explore their conformational space. Comparison with crystal data indicates that relevant conformations are obtained and that, furthermore, the correct biphenyl twist conformation is predicted, as illustrated by the superiority of the resulting model (leave-one-out $q^2 = 0.78$) compared to a random choice of low-energy conformations for each catalyst (average $q^2 = 0.22$). We extend this model by incorporating the MD trajectory directly into a 4D QSSR and by Boltzmann-weighting the contribution of selected minimized conformations, which we refer to as '3.5D' QSSR. The latter method improves on the predictive ability of the 3D QSSR (leave-one-out $q^2 = 0.83$), as confirmed by repeated training/test splits.

INTRODUCTION

The ability to induce chirality in molecules selectively is important in several areas of chemistry, e.g. in natural product synthesis,[1] agrochemicals,[2] and drug design.[3] Therefore asymmetric catalysis is a topic of huge interest, with many advances having been made in recent years.[4] However, there are still many reactions and target molecules for which no effective catalyst is yet known.

Clearly, the ability to produce (and screen) a large number of potential catalysts and a theoretical means by which to guide the search would be a boon to catalyst research. In the field of drug design, the advent of combinatorial chemistry and high-throughput screening has provided an answer to the first problem—and the application of these methods to catalyst design has produced several success stories.[5−7] However, a lesson to learn from the pharmaceutical industry is that maximum gains in efficiency may not be achieved by merely making as many catalysts as possible. Theoretical guidance for the design of catalysts has been lacking, but new computational tools and approaches have begun to appear,[8−13] providing the second vital component for a new age of catalyst discovery. Some techniques, such as stereocartography[10,11] and functional group mapping,[12] are based around molecular modeling, but chemoinformatics tools have also been applied in catalyst discovery, e.g. clustering,[14] docking,[15] database searching,[16] and genetic algorithms,[17] sometimes in combination with molecular modeling.[18]

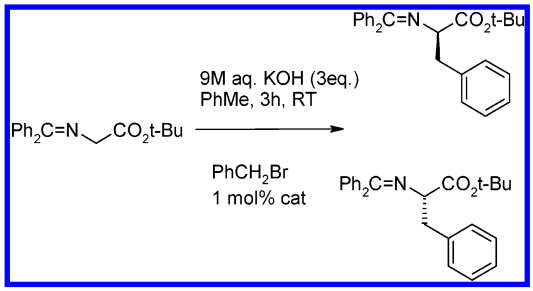
One chemoinformatics-based approach, with a long pedigree in drug design, is Quantitative Structure−Activity Relationships (QSAR), wherein molecular descriptors are correlated with drug activity. When applied to catalyst design, this approach is normally referred to as a Quantitative Structure−Selectivity Relationship (QSSR). Simple correlations between enantiomeric excess (ee) and chirality measures[19] have been reported[20−22] as well as the application of traditional whole-molecule[23] and topological[24] 2D QSAR descriptors; however, given the intrinsic three-dimensional nature of the catalyst design problem, use of descriptors reflecting the three-dimensional structure of molecules may encode more relevant information.[25−27]

In drug design, the most widely used 3D-QSAR method, Comparative Molecular Field Analysis (CoMFA),[28] derives much of its popularity from the ability to allow the user to visualize the regions in 3D space around the molecules where modifications to the electrostatic and steric properties of the molecule will enhance or attenuate the activity of a drug, and clearly this would be a useful technique in catalyst design also. Lipkowitz and Pradhan were the first to apply CoMFA in asymmetric catalyst studies.[29] A related approach using semiempirical quantum chemical methods (QMQSAR)[30] was demonstrated by Kozlowski and co-workers.[31] However, all of these approaches require the construction of the entire molecule, and in the QMQSAR approach, the calculation of transition states. Recently, we showed how CoMFA could be applied in a fragmental approach to produce a predictive 3D QSSR model for computational catalyst design.[32] This requires only the minimization of the substituents, and we have shown that good results can be obtained from this procedure without having to search for transition states.

* Corresponding author phone: +44-115-951-3478; fax: +44-115-951-3562; e-mail: jonathan.hirst@nottingham.ac.uk.
[†] School of Chemistry.
[‡] School of Computer Science & Information Technology.

10.1021/ci050051l CCC: $30.25
Published on Web 05/05/2005

Scheme 1 Reaction Scheme to Which the Catalysts in This Study Are Applied[a]



[a] The alkylation of the glycine imine is enantioselectively controlled by phase-transfer catalysis.

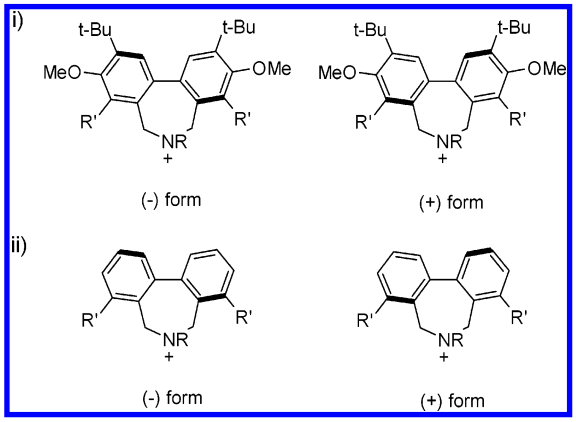


Figure 1. The common structural scaffolds of the phase-transfer catalysts used in this study. Scaffold (i) was used with catalysts **2a**−**8e**, while scaffold (ii) was used only with catalysts **1a**−**e**. Substitution occurs at the points marked NR and R′. The difference between the (+) and (−) conformation of the biphenyl core is denoted by the thick lines on the benzene rings, indicating which edge points out of the page, due to the rotation of the rings.

In this paper, we seek to validate 3D QSSR methods further for catalyst design in conjunction with parallel synthesis, using a new data set consisting of 40 phase-transfer catalysts.[33] However in this case, these catalysts are conformationally labile, necessitating an exploration of the conformational flexibility of these catalysts using molecular dynamics (MD) simulation. We use the results of the MD simulations both to find suitable low-energy conformations for deriving a 3D QSSR and more directly by using the trajectories in a 4D QSSR. We also investigate whether representing each catalyst with a small set of diverse minimized structures can improve the regression modeling. As this approach can be considered a hybrid of the 3D and 4D QSSR methods, we call it a 3.5D QSSR.

METHODS

Data Set. The data set in this study is a library of 40 quaternary ammonium phase catalysts.[33] These are used to promote the asymmetric alkylation of the glycine imine shown in Scheme 1. The catalysts all share the same core structure, based around the biphenyl units indicated in Figure 1. Five catalysts were synthesized using the unsubstituted biphenyl core; all subsequent catalysts were synthesized with the substituted biphenyl core. Structural variation can be introduced by varying the amine (marked as NR) and the substituents of the biphenyl core, at the positions marked as R′. The substituents used to generate the library are given in Figure 2. Structures **1**−**8** were used as substituents on

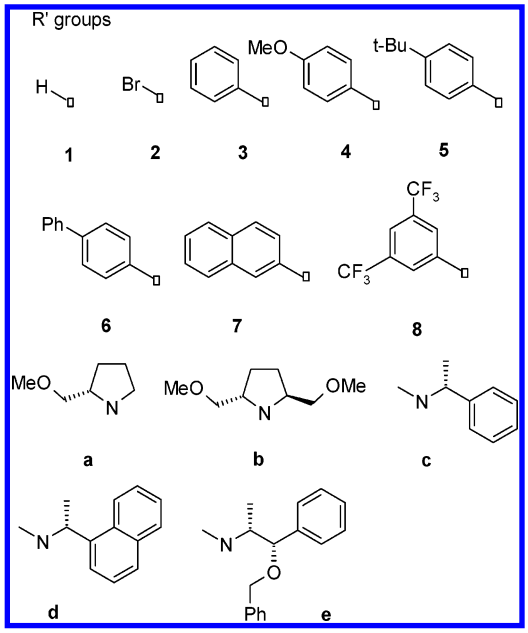


Figure 2. Substituents used in combination with the core structure given in Figure 1. The R′ substituents **1**−**8** were attached to the core at the point marked with a square, and the amines **a**−**e** were used at the NR position. Full enumeration of possible structures led to a library of 40 catalysts.

the biphenyl core, and structures **a**−**e** represent the amines tested. All possible combinations of **1**−**8** and **a**−**e** were used to generate the catalyst library, with the restriction that the same substituent had to be included at both R′ positions. Full enumeration led to 40 catalysts, which will be referred to by the combination of NR and R′ groups as substituents, i.e., **1a**, **2b**, etc. The selectivity of these catalysts was measured by high performance liquid chromatography (HPLC) and the experimentally observed ee's (in favor of the *R* enantiomer) are given in Table 1. For further experimental details, see ref 33. Enantioselectivities range from −30% (i.e. favoring the *S* enantiomer) to 91%. This provides a sufficiently wide range of selectivities to make QSSR modeling feasible.

Conformational Sampling. When considering the structures of the catalysts, the most important feature to consider is the fact that the biphenyl core is fairly flexible and the benzene rings are *not* coplanar, due to the presence of the third ring, containing the amine nitrogen. There are two possible ways for the biphenyl unit to deviate from coplanarity, as shown in Figure 1. To simplify the discussion, we shall refer to the twist form with the top of the left-hand benzene ring pointing out of the page (according to the orientation in Figure 1) as the (−) conformation, and the opposite twist form as the (+) conformation. Whichever of the two forms is the most energetically favorable for each catalyst will have a considerable effect on the positioning of the NR and R′ groups. As the correct alignment of the structures with respect to one another is a key part of the CoMFA methodology we shall be employing, finding the correct biphenyl twist form is vital. Unlike other 3D QSSR catalyst studies, we shall not, however, be concerned with finding the transition state of reactions involving the catalysts. Experiments with topomer CoMFA[34] have shown that finding the binding conformation of ligands in protein binding sites may not be as important as normally thought; finding a *consistent* conformational structure appears to be sufficient.

Table 1. Experimentally Observed Selectivities, with Predicted and Fitted Values Derived Using a 3.5D QSSR Method

R′[a]	NR[a]	ee obs (%)[b]	ee fit (%)[b−d]	ee cv (%)[b,c,e]
1	a	9	0	−5
	b	2	4	1
	c	−2	−3	−4
	d	6	4	1
	e	1	3	2
2	a	4	1	−5
	b	2	6	5
	c	−30	−26	−12
	d	−27	−17	−1
	e	−11	−14	−19
3	a	6	3	2
	b	7	11	12
	c	18	24	29
	d	34	34	35
	e	32	36	34
4	a	0	2	6
	b	6	8	11
	c	21	14	19
	d	36	32	32
	e	29	38	41
5	a	9	10	9
	b	12	19	21
	c	31	30	36
	d	40	45	51
	e	48	48	38
6	a	8	11	13
	b	12	16	17
	c	29	24	30
	d	40	45	49
	e	40	39	33
7	a	20	16	9
	b	40	37	19
	c	42	47	49
	d	78	61	44
	e	40	41	36
8	a	−19	−23	6
	b	8	−7	2
	c	82	84	84
	d	91	89	85
	e	61	71	77

[a] The structures are labeled as in Figure 2. [b] The enantiomeric excesses were converted to a response value for use in regression by $y = \ln([R]/[S])$. [c] 3.5D QSSR values were generated using a three-component PLS model. [d] Fitted values. [e] Predicted values, using leave-one-out cross-validation.

We show below that good results can be obtained with minimized structures from MD simulations and by taking advantage of the constant biphenyl core to provide an obvious alignment rule.

To find a series of potential starting structures, we use the MMFF force field,[35] as implemented in the CHARMM macromolecular package,[36] including a Generalized Born (GB) implicit solvent model.[37,38] Our protocol is to take a series of 'snapshots' at regular intervals and then 'quench' these structures by minimization. While it might be more efficient to use Monte Carlo sampling to obtain our structures, we also wish to use the MD trajectories to generate descriptors for a 4D method, and therefore this makes comparing the two methods easier. The method for structure generation is given in more detail below.

Structure Minimizations. After building, all structures were subjected to a further minimization using the MMFF force field as implemented in CHARMM 28b2. The nonbonded options were set so that 1−4 contributions were scaled by a factor of 0.75, and the electrostatic and van der Waals options were set to TRUNC and VTRUNC, respectively, with the associated cutoff values set sufficiently high so that no distance cutoff was used, and all relevant nonbonded interactions were included in the calculation. The GB solvent model GENBORN was employed, with the dielectric constant set to 2.4, the value for toluene at 298 K. Minimization was carried out using a maximum of 1000 steps of an Adapted Basis Newton−Raphson minimization, followed by up to 1000 steps using the conjugate gradient descent algorithm. The convergence tolerance for early exiting from the minimization was set to 10^{-5}.

Structures for 3D QSAR. Initial MD revealed a substantial kinetic barrier to interconversion between the two twist conformations in all catalysts. Even at 1500 K, very few transitions between the twist forms occur. Therefore, it seemed unlikely that a full exploration of conformational space would take place in a reasonable time with one simulation. However, as we were mainly interested in finding low-energy structures associated with each twist form, two MD simulations were carried out per catalyst, one starting from the (−) form, the other in the (+) form. The GB solvent module was employed, and a time step of 1 fs was applied in all cases. The experimental data were obtained at 298 K, but to increase the conformational space explored, structures were heated to 800 K over 8 ps, with equilibration for a further 5 ps. After equilibration, the production run was carried out for 100 ps. The trajectory was saved every 1 ps. Each conformation from the trajectory was minimized using the same protocol described above. Thus, 200 minimized conformations were produced for each catalyst. The lowest energy conformation was then chosen as the structure adopted by the catalyst for the CoMFA modeling.

The process was repeated for a smaller subset of catalysts at 298 K and 1500 K, with production runs of 1000 ps and extractions of 1000 minimized structures, but no lower energy conformations were found. Therefore, we are confident that the 100 ps runs at 800 K are sampling conformations sufficiently to choose between the twist conformations for each catalyst.

Structures for 4D QSSR. One of the attempts to take into account the dynamic behavior of molecules in a QSAR setting is the 4D QSAR technique originated by Hopfinger and co-workers.[39] In this method, an MD trajectory is used to calculate the time-averaged occupancy of lattice points by various pharmacophoric features in each molecule. We therefore investigated whether the MD trajectories used for generating the structures for the 3D QSSR could be used to extend QSSR modeling into 4D.

As the kinetic barrier to the twist transition was such that transitions between the twist forms during the time scale of the MD runs were either rare or nonexistent, even at 800 K, a second set of MD simulations was performed at 298 K with a starting conformation of the minimum energy structure from the 800 K trajectories. Again, we ran these simulations for 100 ps and sampled 100 conformations, using the GB solvent model with the dielectric value set to 2.4. The resulting structures were *not* minimized. Each structure contributed equally to the descriptor generation (see below).

Structures for 3.5D QSAR. A limitation of the MD trajectories starting from the lowest energy minimum found is that full exploration of the conformation of the catalysts is not achieved. The opposite twist forms may be close in

energy, and hence it is desirable to include these in the model, too. One way to do this is to use a series of minimized structures, a strategy previously attempted by Broughton et al. in QSAR.[40] For this data set, we can take conformations sampled from both of the 800 K MD simulations (an advantage over the 4D QSSR). As this procedure is somewhere between the single conformation used in the 3D QSSR model and the full MD trajectories used in 4D QSSR, we have chosen to call this a 3.5D QSSR model.

Rather than use all 200 structures (many of which may be essentially duplicates of each other), representative conformers were chosen from the 200 by a simple maxmin selection method. The lowest energy structure was chosen to initialize the clustering, and then the root-mean-square deviation (rmsd) of fit was measured for each conformation, excluding hydrogen atoms and any atom in the scaffold. If the rmsd was above a threshold, then the conformer was deemed sufficiently different from the previously chosen conformers to be included in the diverse set. All subsequent conformers were then required to have an rmsd of fit greater than all previously selected conformers in the diverse set. The rmsd threshold was set to 0.9 Å, which provided a good tradeoff between conformational coverage and diversity. This generated between 2 and 32 conformations per catalyst (median: 9.5). Unlike the 4D QSSR method, the chosen conformations do not contribute to the descriptor generation equally but must be Boltzmann weighted (with $T = 298$ K). Initial experiments using an unweighted average were not encouraging, so we only present results using the Boltzmann weighting here. Lukacova and Balaz have criticized the derivation of multiconformational QSARs by linear methods, due to the lack of a physical basis,[41] but approximations to more rigorous conditions have proven useful in other areas of molecular modeling and QSAR: e.g. the assumption that the free energy of binding of a ligand to a receptor can be decomposed into component terms representing the energy of the ligand, receptor, and the complex. Here we hope that the reduction in conformational bias that results from 3.5D and 4D QSAR will offset any deviation from strict LFER principles.

Alignment. Structures were aligned by superimposing the biphenyl core of each catalyst. All 12 aromatic carbon atoms were used for this purpose. The most selective catalyst, **8d**, was the template to which all other structures were fit. Fitting was carried out by means of the least-squares fitting routine in VMD.[42,43] For the 4D QSAR, two alignments steps are necessary. The intercatalyst fitting was carried out as described above. Subsequently, an additional intracatalyst fitting was carried out, where for each catalyst, the 100 snapshot conformations were fit to the corresponding ground-state conformation (used in the 3D QSSR). A similar procedure was carried out for the 3.5D QSSR structures.

Descriptor Generation. As we were interested in the effect of the substituents to the biphenyl core, only substituent atoms contributed to the calculated fields. This has the effect of speeding descriptor calculation and regression, while focusing on the relevant areas of the molecule. A similar scheme was used successfully in our previous work,[32] and models constructed this way were consistently simpler and more predictive than when all the atoms in each catalyst are included.

Descriptors for 3D QSSR. Standard CoMFA parameters[28] were used for descriptor generation, with one exception. The aligned molecules were placed in a 2 Å rectilinear lattice, which extended at least 4 Å from any atom-center in the largest molecule in the data set. Steric descriptors were generated using the Lennard-Jones equation with the Tripos force-field parameters[44] and electrostatic descriptors using the Coulomb potential with Gasteiger−Marsili partial charges.[45] A dielectric constant of 2.4 was used, with an additional distance dependent dielectric term. A 30 kcal mol^{-1} cutoff was applied. While it is normal to apply an interpolation to the fields close to the cutoff value, we have recently shown that use of the 'box' method in CoMFA can give superior results,[46] and therefore we employ this method for this study. In the box method, the descriptor value at each lattice point is the average of the field values sampled at eight points arranged as the corners of a cube centered on the lattice point, with a cube side-length of two-thirds the normal lattice spacing. We refer the reader to ref 46 for further details.

Descriptors for 3.5D and 4D QSSR. The 4D QSSR method as advocated by Hopfinger and co-workers[39] does not use the Lennard-Jones and Coulomb potentials for generating the descriptors. A simple indicator field is used to indicate the presence or absence or certain pharmacophoric types (e.g. positive charge centers, hydrophobes, hydrogen-bond acceptors, etc.) at that grid point. We therefore attempted to use similar rules in the 3.5D QSSR and 4D QSSR models and compared them with the standard Lennard-Jones and Coulombic potentials. For the steric field, we used an indicator field where the grid point was set to one if it was within the van der Waals radius of any substituent atom in the catalyst and zero otherwise. For the electrostatic field, the same procedure was used, but the value assigned to the grid point was set to the partial charge value of the atom. For lattice points within the van der Waals radius of more than one atom, the partial charge assigned was that of the nearest atomic center. Additionally, we explored the use of a simple hydrogen-bond acceptor field (none of the substituents contains hydrogen-bond donors). This was analogous to the steric indicator field, except the only variables set to one were those lattice points within the van der Waals radius of atoms defined as hydrogen-bond acceptors, assigned according to the simple rules using the Tripos atom types used in GASP.[47] Like the 3D QSSR descriptors, the 3.5D and 4D descriptors were calculated for a lattice with 2 Å spacing.

The Response Variable. While some authors have derived QSSRs using the ee's directly as the response variable,[24] we transform the response such that it corresponds to a linear free-energy relationship, $\Delta G = -RT \ln K$. We transform the ee to an equilibrium constant by using $K = [R]/[S]$, i.e., the ratio of the proportion of the *R* to the *S* enantiomer. We then take the log of this value to reach a linear free energy relationship form. This response, *y*, is then correlated with the descriptors generated as described above using Partial Least Squares (PLS) regression,[48] implementing the SIMPLS algorithm of de Jong[49] with an in-house program written in C++.

Scaling and Filtering. Block scaling (COMFA_STD) was employed and to speed computation the least varying 80% of the descriptors in each block were removed. This never

worsened results compared to keeping all variables, and in many cases gave a slight improvement in model predictivity and parsimony.

Cross-Validation. To assess the predictivity of the models and to choose the correct dimensionality of the PLS model, we employ leave-one-out (LOO) cross-validation.[50] It is normal to choose the model that minimizes the value of s_{PRESS}

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

where A is the number of components extracted, N is the number of observations (in this case, catalysts), and the predicted residual sum of squares (PRESS) is given by

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

where y_n is the observed selectivity of catalyst n and $\hat{y}_n$ is the corresponding predicted value.

While minimizing s_{PRESS} penalizes more complex models (i.e. those with a larger number of components), this is not always adequate. As an alternative, we instead set a minimum threshold to the LOO q^2 increase with successive components, where q^2 is defined as

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

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

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

where $\bar{y}$ is the average of the observed selectivity response values. Allowing an extra component with an increase of q^2 of 0.01 was likely only to be fitting noise, while an increase of 0.05 units was too strict. We therefore adopted 0.025 units as a good threshold. Kellogg and co-workers have also used this criterion in CoMFA modeling.[51] This model selection method normally produces identical models to those that minimize s_{PRESS}, but in situations where they differ it results in more parsimonious models. As a result, the q^2 values are often slightly smaller than those achieved when minimizing s_{PRESS}, so we shall consider both q^2 and the coefficient of determination, R^2 (determined analogously to the q^2, but using the fitted, not predicted selectivities) when assessing the predictive quality of the model. While a high q^2 is necessary for a predictive model, the corresponding R^2 should not be larger by more than 0.2 units.

RESULTS

3D QSSR. Using the MD protocol to select a single conformation per catalyst, combined with the field generation and cross-validation procedure outlined above, we obtained a four-component model with LOO $q^2 = 0.78$ and $R^2 = 0.94$. Increasing the grid resolution to 1 Å gave no improvement. The q^2 value indicates a model with good predictive power, and the discrepancy between q^2 and R^2 values is not too large, indicating that we can be fairly confident that no major overfitting is occurring. It is gratifying that the MD protocol is capable of producing a 3D QSSR model with good statistics, but the possibility remains that the procedure was unnecessary: could an arbitrary choice of twist conformation lead to equally good results?

Table 2. Results of Different Conformation Selections on Resulting 3D QSSR Models

method	q^2	R^2	ONC[a]
MD[b]	0.78	0.94	4
all (+)[c]	0.34	0.68	2
all (−)[c]	0.28	0.54	1
reverse[d]	0.65	0.88	3
random − mean[e]	0.22	0.59	1.3
random − best[f]	0.70	0.94	4

[a] Optimum number of components, chosen by LOO cross-validation. [b] Conformations chosen as the lowest energy minimized structure from 200 trajectories over 200 ps of MD simulation. [c] All catalysts were represented by the lowest energy structures corresponding to either the (+) or (−) conformations as marked in Figure 1. [d] Catalysts were represented by the lowest energy conformation corresponding to the other twist form to that predicted to be the lowest in energy by the MD simulations. [e] The minimum energy (+) or (−) twist form was chosen at random for each catalyst; results are reported as an average over 100 separate runs. [f] The best result found from the 100 random assignments of the (+) or (−) twist form.

We have assumed that a suitable conformation for the QSSR would be the lowest energy structure found during two MD simulations, the first starting from the (+) twist conformation, the second from the opposite conformation. To test the validity of this assumption, we repeated the CoMFA several times, using different combinations of the lowest (+) or (−) conformation found for each catalyst. Three alternative conformational hypotheses are of interest: (i) an alignment using all the catalysts in the (+) conformation; (ii) an alignment using all the catalysts in the (−) conformation; and (iii) an alignment which reverses the conformations chosen on the basis of the MD simulations. The first two hypotheses test whether it is necessary to use the information from the MD simulations in choosing a conformation, or whether it is sufficient to choose a single twist conformation and apply that to each catalyst. If it is necessary to use different twist conformations, the third hypothesis tests whether the relative energies of the minimized conformations are effective in predicting which is the correct conformation to use—if equally good models can be obtained from this hypothesis as from the MD-based conformational hypothesis, then we can only conclude that *relative* conformations are necessary, that is, some catalysts should take up the opposite twist conformation to others, but details of the *absolute* conformations cannot be elucidated by this method.

Table 2 summarizes the R^2, optimum number of components (ONC), and LOO q^2 of the three alternative hypotheses. The MD-based QSSR is given in the top row of the table as a reference. The models based on only one twist conformation result in inferior models, with resulting q^2 values of 0.34 for the 'all (+)' conformation and only 0.28 for the 'all (−)' model. Therefore, it seems that catalysts are adopting different twist conformations, and this needs to be reflected by the QSAR model. The 'reverse MD' model utilizes the information from the MD simulation but assigns the opposite twist conformation to that predicted by a comparison of the minimized energies. The resulting q^2 (0.65) is superior to the single-conformation models but still substantially worse

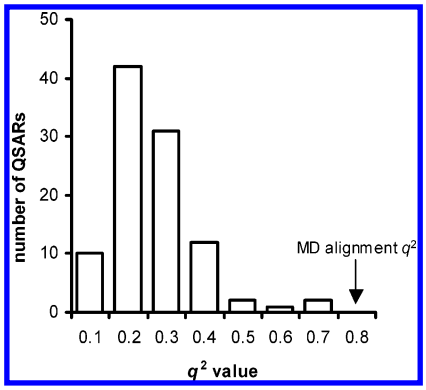


Figure 3. A histogram of the q^2 distribution for the alignment scramble test. The labeling of the *x*-axis indicates that the QSARs in that bin attained a value greater than the value in the previous bin and less than the number indicated. The bin into which the q^2 obtained using the MD protocol falls is indicated also.

than that using the conformations predicted to be the lowest energy. This shows that not only is it necessary to take into account differences in twist conformations in this model but that the MD simulation does accurately reflect the conformational energetics, which in turn improves the resulting QSSR model.

Despite these encouraging results, many other conformational hypotheses are possible, by choosing one of the (+) or (−) twist conformations at random for each catalyst and then building a 3D QSSR based on them. To test whether a random choice is better than the conformations predicted by the MD simulation, a variant on the 'scramble set' test was carried out. Instead of scrambling the predicted selectivities, the choice of (+) or (−) conformation was chosen at random for each catalyst and the q^2 of the resulting model recorded. This was repeated 100 times, and the distribution of the resulting q^2 values is shown in Figure 3. The average q^2 of 100 runs was 0.22, with a maximum value of 0.70. 95% of q^2 values obtained were lower than 0.40. Thus, in general, the result of choosing either the (+) or (−) twist conformations at random was substantially worse than (and, at best, still inferior to) the method we followed. Hence, we conclude that choosing the conformation with the lowest energy from two short MD simulations per catalyst finds a chemically relevant set of structures for analysis.

Effect of Solvent and Dielectric. To assess the importance of the GB solvent model, we repeated the MD simulations without the GB option and derived a 3D QSSR model with the lowest energy conformations found. This resulted in a four-component model with $q^2 = 0.69$, $R^2 = 0.90$. This is a more complex model, which is perhaps overfit. This suggests that the GB model samples a more relevant set of conformations for a 3D QSSR analysis. One other consideration is that, due to the phase-transfer behavior of the system, it is possible the catalysis takes place in a more aqueous environment than reflected in the use of the pure toluene dielectric. Additionally, GB solvent models are normally parametrized to reproduce the Poisson−Boltzmann results with a large exterior dielectric (usually 80). We therefore repeated the MD simulation with the dielectric constant set to 80 and used this value for the electrostatic CoMFA field generation. The resulting QSSR model consisted of four components with $q^2 = 0.77$ and $R^2 = 0.94$, almost identical to the values obtained when using the much lower dielectric value for

Table 3. Summary of the 3−4D QSSR Models Investigated

QSSR method	LOO q^2	R^2	ONC[a]
3D CoMFA	0.78	0.94	4
3D in vacuo[b]	0.69	0.90	4
3D indicator[c]	0.70	0.85	2
3.5D CoMFA	0.73	0.97	5
3.5D indicator[c]	0.79	0.93	3
3.5D indicator[c] + HBA[d]	0.82	0.95	3
4D CoMFA	0.71	0.96	4
4D indicator[c]	0.76	0.86	2

[a] Optimal Number of Components. [b] Structures generated with MD simulation in vacuo. [c] Descriptors generated with indicator values. [d] Includes the hydrogen-bond acceptor field.

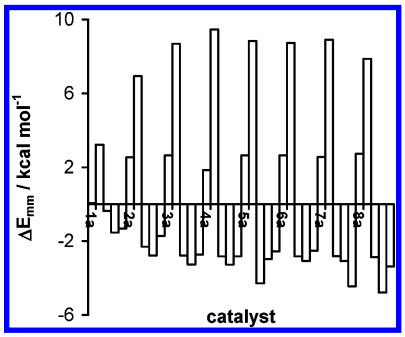


Figure 4. The difference in energies between the lowest energy (+) twist conformation and the lowest energy (−) twist conformation found for each catalyst using the MD simulation with GB solvent model. A negative value indicates that the (+) conformation is more favorable.

toluene. It therefore seems that the extra time spent using the GB solvent model is worthwhile, and results are not sensitive to the choice of dielectric constant. While the computational time required to generate the conformers approximately doubles (between 4 and 20 min on a 2.8 GHz Xeon with 1 GB RAM were required to generate a MD trajectory for each catalyst) with the GB solvent model, it is still sufficiently brief that a large number of structures can be processed in a reasonable time. The results of our 3D QSSR modeling are summarized in Table 3.

Origins of the Energy Difference. The difference in energy between the lowest energy (−) and (+) twist form found for each catalyst is shown in Figure 4. Catalysts are grouped by amine and then the phenyl substituent, R′. The obvious periodicity in the figure suggests that the amine affects the twist of the molecule most substantially, with the nature of the phenyl substituents having a much smaller effect. Using either of the two cyclic amines at NR always promotes the (−) twist over the (+) form, while the (+) form is favored with all other amines. It is clear that the cyclic amine **b** promotes the stability of the (−) conformation to a much larger extent than using **a**. The origin of these differences can be rationalized by an inspection of the amine structures. All the acyclic amines have the same ethyl and methyl substitution pattern. For the (−) conformations, there is a larger steric hindrance between the ethyl substituent and one of the benzene substituent R′ groups. To relieve this steric contact, the R′ group is rotated away somewhat from its optimal position leading to less favorable dihedral and angle strain values. Hence, the (−) conformations are destabilized with respect to the (+) conformation. The opposite holds true with the cyclic amines. The cyclic form of the amine causes the methoxy group (the one shown

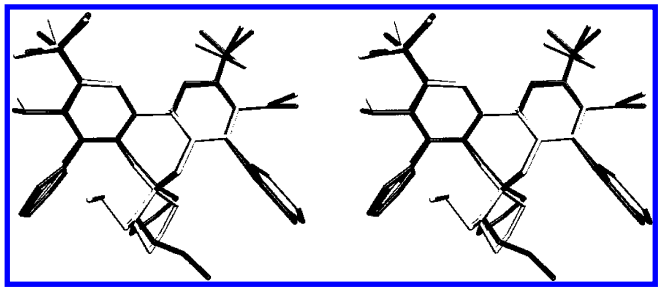


Figure 5. A stereoview comparison of the X-ray structure of catalyst **3a** (dark color) and the lowest energy structure found using an MD simulation with GB implicit solvent (light color), with hydrogens removed for clarity. Structures were superimposed using the carbon atoms of the biphenyl core.

pointing into the page in **a** and **b**) to point toward the R′ groups to a greater extent in the (+) twist form, resulting in the same destabilization as was observed for the acyclic amines, but this time favoring the (−) form. The energy difference is rather low for some conformations; given the inherent uncertainty in the energies from the force field, the choice of which twist form to use is not necessarily clear-cut. We investigate this further in the next section and present a possible solution (a multiconformational, 3.5D QSSR).

Comparison with X-ray Structure. The crystal structure of catalyst **3a** was determined, and so this provides a test of whether the MMFF force field was capable of producing a likely conformation for these structures. Figure 5 shows the lowest energy minimum of **3a** found by the MD protocol, superimposed over the X-ray structure, which has not been minimized, to make the comparison as strict as possible—however, no major conformational change was observed during an exploratory minimization of the X-ray structure, with a rmsd between the minimized and unminimized structure of 0.6 Å. The structures were superimposed using the biphenyl rings with the least-squares atom fit procedure in SYBYL 6.91.[52] From the overlay, it can be seen that the main difference in conformation arises from a difference in the amine ring conformation, which causes the methoxy group to point in different directions. It should be borne in mind that we seek to produce a set of consistent ground-state structures with the correct twist forms, rather than to reproduce the crystal structure of the crystal. Nonetheless, the rest of the molecule shows a good overlap, with a total heavy-atom rmsd of 1.6 Å. The X-ray structure biphenyl angle is −46°, while that of the MD-derived structure is −56°, showing that this method predicts the correct twist form for this catalyst. Comparison of the minimum structure of **8a** with its X-ray structure (data not shown) shows a similar quality of fit, with the same correct prediction of the twist direction. Details of the crystal structure are available in the Supporting Information.

3.5D QSSR. To overcome any bias introduced by the choice of a single conformation, particularly in the case where there was little difference between the twist forms of the catalysts, we implemented a 3.5D QSSR method, with Boltzmann weighting of the contribution of each of the chosen conformations. For most catalysts, only the lowest (+) and (−) conformations contributed significantly to the final descriptors. Using the CoMFA fields yielded a five-component model with $q^2 = 0.73$ and $R^2 = 0.97$. Once again, increasing the lattice resolution to 1 Å did not improve the model. It seems that, in this case, at least, the extra conformations only add more noise to the model, as indicated by the extra component that was extracted but with a decrease in predictive ability.

As indicator fields have been used successfully in 4D-QSAR applications, we investigated whether they represent a suitable level of detail for use in a 3.5D QSSR. To evaluate how much (if any) improvement to the model can be achieved by the use of multiple conformations, we carried out a 3D QSSR (i.e. one conformation per catalyst) using the indicator fields. A two-component model resulted, with $q^2 = 0.70$ and $R^2 = 0.85$, which is more parsimonious than the CoMFA 3D QSSR model, but less predictive. However, including the multiple conformations to generate a 3.5D QSSR model yielded an improved three-component model with $q^2 = 0.79$ and $R^2 = 0.93$. This is both more parsimonious than the 3D QSSR CoMFA model, more predictive, and less overfit.

We also investigated whether the addition of a hydrogen-bond acceptor indicator field could improve the model further. Addition of this third field led to a three-component model with $q^2 = 0.82$ and $R^2 = 0.95$, a further improvement. These results indicate that the use of multiple conformations in QSSR can yield more predictive models, but the choice of descriptor is crucial—the more accurate Lennard-Jones and Coulomb potentials used in CoMFA add too much noise. Furthermore, there appears to be a slight improvement to using a hydrogen-bond acceptor indicator field in addition to the steric and electrostatic fields.

4D QSSR. For the full 4D QSSR models we followed a similar protocol to that given above for the 3.5D QSSR. We began by generating standard CoMFA fields for each of the 100 conformations generated during the 100 ps MD run at 298 K with the GB solvent model for each catalyst. Each of the fields was then averaged and used in the PLS regression as before. As with the 3.5D QSSR, however, the resulting four-component model was inferior to the single conformation model with $q^2 = 0.71$ and $R^2 = 0.96$. Again, this suggests that any extra 'signal' that results from adding extra conformations in the manner described above is swamped by the extra 'noise' that is introduced.

However, like the 3.5D QSSR model, better results can be obtained with the indicator fields. Use of the steric and electrostatic fields gave a two-component model with $q^2 = 0.76$ and $R^2 = 0.86$. This is slightly less predictive than the 3D QSSR model but more parsimonious and the least overfit of all the models derived so far. This model might therefore be of interest as the one in which we have the most confidence in using on an external test set. Again, there is a possibility of using the extra hydrogen-bond acceptor field. However, its use gave no improvement in this situation, yielding a two-component model with $q^2 = 0.75$, $R^2 = 0.85$ close, but inferior, to the steric + electrostatic 4D QSSR.

That the 4D model is less predictive than the 3.5D model may be explained by the fact that the 3.5D model used structures with both twist conformations, whereas only the trajectories starting from the most stable twist conformations were used for the 4D model. At 298 K, none of the catalysts made the transition from one twist conformation to the other during the 100 ps of simulation, so full sampling of the conformations was not carried out. The results of the 3.5D and 4D QSSR modeling are summarized in Table 3.

Table 4. Results of Two-Deep Cross-Validation Averaged over 100 Splits[a]

model	q^2
3D	0.64 ± 0.14
3.5D	0.70 ± 0.12
3.5D+HBA[b]	0.73 ± 0.10
4D	0.71 ± 0.09

[a] Model was generated by 4-fold cross-validation to generate a series of 30/10 training/test splits. Within the training set, LOO cross-validation was used to determine the model dimensionality. The reported q^2 is based on the predictions of the 10 member test sets, over all 4-folds. The whole process was then repeated 100 times. [b] Includes the hydrogen-bond acceptor indicator field.

Two-Deep Cross-Validation. Based on the LOO results, there are four models of similar predictive value: the 3D QSSR model using the CoMFA-style fields, a 3.5D model using indicator fields, the 3.5D model with the addition of the hydrogen-bond acceptor indicator field, and the 4D QSSR model with steric and electrostatic indicator fields. Five-fold cross-validation, repeated 100 times gave similar results to LOO cross-validation, with the expected reduced q^2, but the same ordering to the predictive quality of the models. Hence, we do not show these results. A sterner test of the models' predictivities would be the use of an external test set, not involved in the model building. Unfortunately, there are not enough compounds in this data set to enable the creation of a test set of sufficient size for the results to be more reliable than cross-validation (approximately 20 would be needed, with approximately 30 molecules left in the training set),[53,54] so instead we employ two-deep cross-validation,[55] as recommended for PLS by Jonathan and co-workers.[56] It works as follows: we split the 40 catalysts into four groups of 10 catalysts each, as in 4-fold cross-validation. We then form a training set with three of the groups and build a model using these 30 catalysts. Model selection is carried out using LOO cross-validation employing *only* the 30 training catalysts. The 10 remaining molecules left are then used as an independent test set. This process is repeated, using one of the other groups of 10 catalysts as the test set, until all molecules have been predicted, just as in normal cross-validation. The test set q^2 is calculated as for regular cross-validation, using all 40 predicted values, except that the mean of the selectivities used to calculate the Total Sum of the Squares is the value of the corresponding training set selectivity mean. This was done to make the calculation more similar to a normal test set q^2. The entire two-deep cross-validation is carried out 100 times, and the mean values are reported in Table 4. The 3.5D and 4D QSSR models perform similarly, with the 3.5D model with the hydrogen-bonding acceptor field again the best. It would appear that there is a small but consistent increase in information in using this descriptor in the 3.5D model. The 3D QSSR model performs noticeably less well than the others, although the result is not a poor one by any means. Nonetheless, this falloff in performance may be a result of slight overfitting, as it was the most complex model with four components. From this test, there is evidence that 3.5D and 4D QSSR methods are simpler, yet more predictive than the 3D QSSR models. The overall best model appears to be the 3.5D QSSR model with the addition of the hydrogen-bond acceptor indicator field, so this is the one we shall study in further detail.

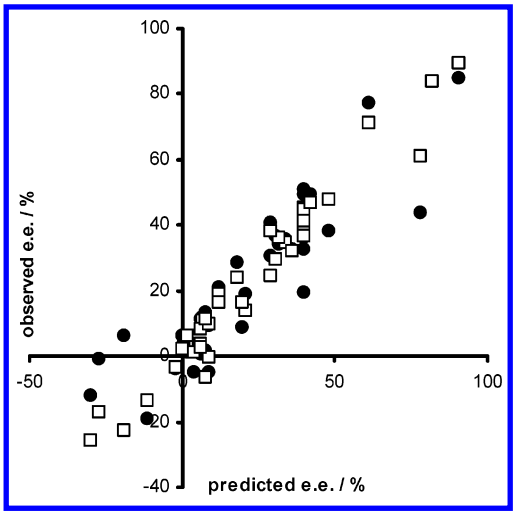


Figure 6. Plot of observed experimental selectivities against those resulting from a three-component 3.5D QSSR model, including hydrogen bonding indicator field. Squares indicate the fitted values, filled circles the LOO cross-validated values.

Predicted against Observed Values. A plot of the predicted against the observed ee's for the 3.5D model including the hydrogen-bond acceptor indicator field is shown in Figure 6. Both the fitted and the cross-validated results are shown. Values used to make this plot are tabulated in Table 1. The R^2 and q^2 of 0.95 and 0.83, respectively, corresponds to an rms error of prediction of 11% in prediction and 6% in fitting. The contribution of the steric and electrostatic fields is more or less equal (35% and 38%, respectively) with the hydrogen bond acceptor being slightly less important (26%). Overall, the predictions are satisfactory for most compounds. There are two main groups of compounds for which the fit is perhaps less good: the compounds predicted to have the opposite enantioselectivity to the majority of the catalysts have the direction of their selectivity wrongly predicted in some cases. However, as the ee's are modest for these catalysts, this is not a major concern. The other two noticeable outliers are compounds **7b** and **7d**, the selectivity of both being underpredicted by ~30%. Compound **7b** is the only catalyst containing the cyclic amine that has appreciable selectivity, and this may be why the model underpredicts its performance. The underprediction for **7d** is perhaps less clear. Although catalysts containing amine **d** also produce generally poor catalysts, the best catalyst (**8d**) also contains this amine. Isolating the causes for this misprediction is a subject of future work, along with the use of chemometric techniques for training set selection.

Visualization. We visualize the results by mapping the regression coefficients back onto the corresponding lattice points. Coefficients were standardized by multiplication with the standard deviation of the corresponding variable. Due to the large number of lattice points we have carried out a simple clustering procedure to simplify the visualization. First we calculated the mean absolute value of the coefficients. Grid points with coefficients of less than this value were removed. Local maxima of the regression coefficients were identified by comparing each of the coefficients with their 26 nearest neighbors (i.e. adjacent lattice points in any or all of the *x*, *y*, or *z* directions). Comparisons were considered invalid and ignored if the coefficient was of the opposite

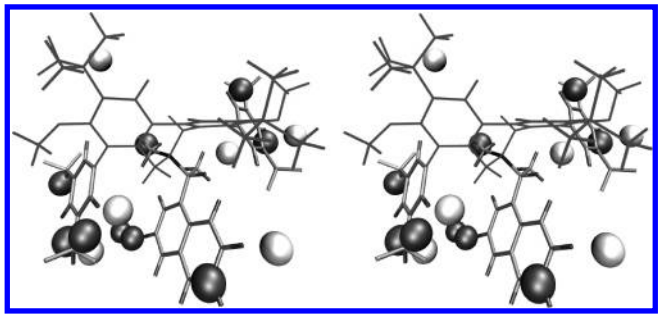


Figure 7. Stereoview of important steric lattice positions for the 3.5D QSSR model. The dark spheres indicate regions where steric bulk increases enantioselectivity, and white spheres indicate those regions where steric bulk reduces selectivity.

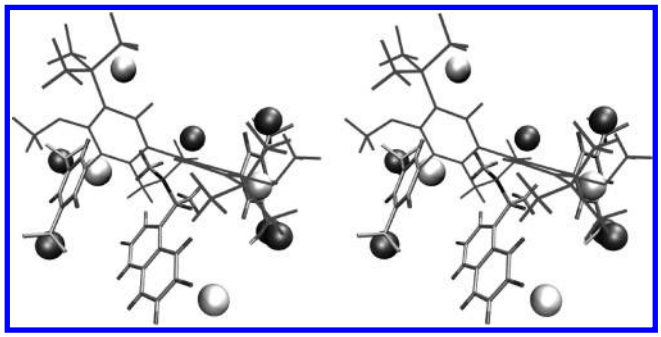


Figure 8. Stereoview of important hydrogen-bond acceptor lattice positions for the 3.5D QSSR model. The dark spheres indicate regions where the presence of hydrogen-bond acceptors increase selectivity, and light spheres indicate regions where hydrogen-bond acceptors reduce selectivity.

sign or the lattice point had been filtered due to low variance prior to the regression. Additional criteria were applied in order for the point to qualify as a maximum: the absolute value of the coefficient had to be larger than the mean absolute value of the coefficients and to have at least one neighboring point for which a valid comparison could be made. This removes lattice points that do not contribute significantly to the model and 'singleton' lattice points, which had no neighbors with coefficients of the same sign or which had been filtered, for which interpretation is difficult.

Figures 7 and 8 show the position of the most important grid points as identified by the clustering procedure given above for the 3.5D model. Qualitatively similar results were obtained with the 3D and 4D QSAR models, so these are not displayed, although it is reassuring that all methods allow for a similar interpretation. The most selective catalyst, **8d**, is also displayed in its lowest energy conformation. The R′ and NR substituents are shown in thicker lines than the biphenyl core, which played no role in the descriptor generation. In the steric field, shown in Figure 7, the dark spheres represent lattice points where steric bulk increases selectivity and the white spheres regions where steric bulk decreases selectivity. Important selectivity-enhancing lattice points are found at CF_3 of the R′ groups. This reflects the fact that only the substituents **7** and **8** are substituted in either of the meta positions, and most of the selective catalysts have one of these substitutions. A further selectivity enhancing grid point is shown in the region of the naphthalene ring of the amine—the first and third most selective catalysts contain this amine, and the other catalysts do not probe this region of space. Lattice points where steric occupation results in selectivity reduction are clearly shown in the two corners of the plot where there is no steric occupation by **8d**—these are positions where the R′ groups would be if the (−) twist form predominated. This highlights the lack of selectivity of the catalysts that possess a cyclic amine, which stabilizes that form over the (+) twist. The electrostatic plot (data not shown) is less informative. Examination of the unclustered lattice points shows very few separated contiguous regions, making interpretation difficult, so it is not a problem with the lattice point clustering procedure itself. A similar picture is evident in the 3D fields, so the difficulty does not lie with the use of multiple conformations or indicator fields. The presence of a negative charge lattice point near one of the electronegative CF_3 groups again indicates the selectivity enhancing properties of that substituent. However, for other regions, these are paired with positive charge lattice points, so this plot can be used to design new catalysts only with difficulty. The hydrogen-bond acceptor field (Figure 8) mirrors the steric field, although it has the benefit of being rather clearer than the steric field, due to a smaller number of nonzero variance lattice points. The importance of the CF_3 groups is even clearer in this plot. As more than one conformation is used to generate these descriptors, displaying more than one conformation may help. However, for the data set, experimentation with such a protocol did not help in interpretation. Clearly, this is an important area for future research.

CONCLUSIONS

Previously, we have successfully applied CoMFA modeling to a set of phase-transfer catalysts. In this study, we have further confirmed the validity of this approach with a new combinatorial library of catalysts. We have extended the technique by combining it with GB MD simulations to produce a range of likely starting conformations and also relative energies. This is particularly vital for this data set; despite the common scaffold that all the molecules share, the two twist conformations can produce very different arrangements of substituents, and a naïve choice of conformations (i.e. assuming all catalysts twist the same way) produces a very poor model. This may have implications for systems for producing consistent conformations, e.g. topomer generation.[34]

We have also shown that catalysts modeling can be improved further by going beyond 3D QSSR, by either averaging over the fields produced by an MD trajectory (4D QSSR), or from Boltzmann weighting a set of diverse minimized conformers. In these cases, it seems necessary to use simple indicator fields rather than those produced by a Lennard-Jones or Coulombic potential, to avoid overwhelming the model with noise. Our approach differs slightly from those presented by other workers. Note that, unlike the 4D QSAR method introduced by Hopfinger and co-workers,[39] we have found acceptable models without the need for optimization by genetic algorithms, which in some cases appears to overfit the data, compared to CoMFA.[57] Also, unlike the observations of Broughton and co-workers,[40] when we averaged over minimized structures, a Boltzmann weighting produced the best results, rather than unweighted averaging. Apart from its use in catalyst design, the results presented have significance for the practice of 3D and 4D QSAR in general.

Areas for improvement in this method involve using other chemoinformatics techniques to select a balanced training set, and embedding the QSSR modeling into an iterative system of testing and predicting. A Monte Carlo method rather than MD could generate the conformations for the Boltzmann-weighted models. The component-based nature of the combinatorial library could also be exploited to generate a series of conformations of the substituents and the scaffold separately, which could then be combined, similar to the CONAN method.[58] Also, variable selection to reduce the number of lattice points in the model may be useful. Research is currently underway in all these areas.

ACKNOWLEDGMENT

We thank the EPSRC for support (GR/S75765/01) and an equipment grant (GR/R62052/01) for computers. We thank Tim Watson and Matt Wood for IT assistance.

Supporting Information Available: Minimized structures used for 3D and 3.5D structures, aligned and with Gasteiger−Marsili partial charges, in mol2 format, crystallographic data for catalyst **3a** in CIF and PDF format. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ohshima, T. Enantioselective total syntheses of several bioactive natural products based on the development of practical asymmetric catalysis. *Chem. Pharm. Bull.* **2004**, *52*, 1031−1052.
(2) Blaser, H. U.; Spindler, F. Enantioselective catalysis for agrochemicals. The case histories of (S)-metolachlor, (R)-metalaxyl and clozylacon. *Top. Catal.* **1997**, *4*, 275−282.
(3) Beroza, P.; Suto, M. J. Designing chiral libraries for drug discovery. *Drug Discovery Today* **2000**, *5*, 364−372.
(4) Jacobsen, E. N.; Pfaltz, A.; Yamamoto, H. *Comprehensive Asymmetric Catalysis*; Springer-Verlag: New York, 1999.
(5) Murphy, V.; Volpe, A. F., Jr.; Weinberg, W. H. High-throughput approaches to catalyst discovery. *Curr. Opin. Chem. Biol.* **2003**, *7*, 427−433.
(6) Gennari, C.; Piarulli, U. Combinatorial libraries of chiral ligands for enantioselective catalysis. *Chem. Rev.* **2003**, *103*, 3071−3100.
(7) Hechavarría Fonseca, M.; List, B. Combinatorial chemistry and high-throughput screening for the discovery of organocatalysts. *Curr. Opin. Chem. Biol.* **2004**, *8*, 319−326.
(8) Cundari, T. R.; Deng, J.; Fu, W.; Klinckman, T. R.; Yoshikawa, A. Molecular modeling of catalysts and catalytic reactions. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 941−948.
(9) Lipkowitz, K. B.; Kozlowski, M. C. Understanding stereoinduction in catalysis via computer: new tools for asymmetric synthesis. *Synlett* **2003**, 1547−1565.
(10) Lipkowitz, K. B.; D'Hue, C. A.; Sakamoto, T.; Stack, J. N. Stereocartography: a computational mapping technique that can locate regions of maximum stereoinduction around chiral catalysts. *J. Am. Chem. Soc.* **2002**, *124*, 14255−14267.
(11) Lipkowitz, K. B.; Sakamoto, T.; Stack, J. Using stereocartography for predicting efficacy of stereoinduction by chiral catalysts. *Chirality* **2003**, *15*, 759−765.
(12) Kozlowski, M. C.; Panda, M. Computer-aided design of chiral ligands. Part 2. Functionality mapping as a method to identify stereocontrol elements for asymmetric reactions. *J. Org. Chem.* **2003**, *68*, 2061−2076.
(13) Shimizu, H.; Ishizaki, T.; Fujiwara, T.; Saito, T. A novel approach for investigating enantioselectivity in asymmetric hydrogenation. *Tetrahedron: Asymmetry* **2004**, *15*, 2169−2172.
(14) Cundari, T. R.; Deng, J.; Pop, H. F.; Sârbu, C. Structural analysis of transition metal β-X substituent interactions. Towards the use of soft computing methods for catalyst modeling. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1052−1061.
(15) Harriman, D. J.; Deslongchamps, G. Reverse-docking as a computational tool for the study of asymmetric organocatalysis. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 303−308.
(16) Kozlowksi, M. C.; Panda, M. Computer-aided design of chiral ligands Part I. Database search methods to identify ligand types for asymmetric reactions. *J. Mol. Graph. Modell.* **2002**, *20*, 399−409.
(17) Cawse, J. N.; Baerns, M.; Holena, M. Efficient discovery of nonlinear dependencies in a combinatorial catalyst data set. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 143−146.
(18) Kozlowski, M. C.; Waters, S. P.; Skudlarek, J. W.; Evans, C. A. Computer-aided design of chiral ligands. Part III. A novel ligand for asymmetric allylation designed using computational techniques. *Org. Lett.* **2002**, *4*, 4391−4393.
(19) Zabrodsky, H.; Avnir, D. Continuous symmetry measures. 4. Chirality. *J. Am. Chem. Soc.* **1995**, *117*, 462−473.
(20) Gao, D.; Schefzick, S.; Lipkowitz, K. B. Relationship between chirality content and stereoinduction: identification of a chirapore. *J. Am. Chem. Soc.* **1999**, *121*, 9481−9482.
(21) Lipkowitz, K. B.; Schefzick, S.; Avnir, D. Enhancement of enantiomeric excess by ligand distortion. *J. Am. Chem. Soc.* **2001**, *123*, 6710−6711.
(22) Alvarez, S.; Schefzick, S.; Lipkowitz, K.; Avnir, D. Quantitative chirality analysis of molecular subunits of bis(oxazoline)copper(II) complexes in relation to their enantioselective catalytic activity. *Chem. Eur. J.* **2003**, *9*, 5832−5837.
(23) Klanner, C.; Farrusseng, D.; Baumes, L.; Lengliz, M.; Mirodatos, C.; Schüth, F. The development of descriptors for solids: teaching "catalytic intuition" to a computer. *Angew. Chem., Int. Ed.* **2004**, *43*, 5347−5349.
(24) Jiang, C.; Li, Y.; Tian, Q.; You, T. QSAR study of catalytic asymmetric reactions with topological indices. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1876−1881.
(25) Aires-de-Sousa, J. Gasteiger, J. New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 369−375.
(26) Hoogenraad, M.; Klaus, G. M.; Elders, N.; Hooijschuur, S. M.; McKay, B.; Smith, A. A.; Damen, E. W. P. Oxazaborolidine mediated asymmetric ketone reduction: prediction of enantiomeric excess based on catalyst structure. *Tetrahedron: Asymmetry* **2004**, *15*, 519−523.
(27) Funar-Timofei, S.; Suzuki, T.; Paier, J. A.; Steinreiber, A.; Faber, K.; Fabian, W. M. F. Quantitative structure−activity relationships for the enantioselectivity of oxirane ring-opening catalyzed by epoxide hydrolases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 934−940.
(28) 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*, 5959−5967.
(29) Lipkowitz, K. B.; Pradhan, M. Computational studies of chiral catalysts: a comparative molecular field analysis of an asymmetric Diels−Alder reaction with catalysts containing bisoxazoline or phosphinooxazoline ligands. *J. Org. Chem.* **2003**, *68*, 4648−4656.
(30) Dixon, S.; Merz, K. M., Jr.; Lauri, G.; Ianni, J. C. QMQSAR: utilization of a semiempirical probe potential in a field-based QSAR method. *J. Comput. Chem.* **2005**, *26*, 23−34.
(31) Kozlowski, M. C.; Dixon, S. L.; Panda, M.; Lauri, G. Quantum mechanical models correlating structure with selectivity: predicting the enantioselectivity of β-amino alcohol catalysts in aldehyde alkylation. *J. Am. Chem. Soc.* **2003**, *125*, 6614−6615.
(32) Melville, J. L.; Andrews, B. I.; Lygo B.; Hirst J. D. Computational screening of combinatorial catalyst libraries. *Chem. Commun.* **2004**, 1410−1411.
(33) Lygo, B.; Allbutt, B.; James, S. R. Identification of a highly effective asymmetric phase-transfer catalyst derived from α-methylnaphthylamine. *Tetrahedron Lett.* **2003**, 44, 5629−5632.
(34) Cramer, R. D. Topomer CoMFA: A design methodology for rapid lead optimization. *J. Med. Chem.* **2003**, *46*, 374−388.
(35) Halgren, T. A. Merck molecular force field. 1. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(36) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(37) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(38) Dominy, B. N.; Brooks, C. L., III Development of a Generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **1999**, *103*, 3765−3773.
(39) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; 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.
(40) Broughton, H. B.; Gordaliza, M.; Castro, M.-A.; Miguel del Corral, J. M.; San Feliciano, A. modified CoMFA methods for the analysis of antineoplastic effects of lignan analogues. *J. Mol. Struct. (THEOCHEM)* **2000**, *504*, 287−294.

(41) Lukacova, V.; Balaz, S. Multimode binding in receptor site modeling: implementation in CoMFA. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2093-2105.
(42) Humphrey, W.; Dalke, A.; Schulten, K. VMD − Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33−38.
(43) Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1978**, *A34*, 827−828.
(44) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982−1012.
(45) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(46) Melville, J. L.; Hirst, J. D. On the stability of CoMFA models. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1294−1300.
(47) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.
(48) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(49) de Jong, S. SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251−263.
(50) Wold, S. Cross-validatory estimation of number of components in factor and principal components models. *Technometrics* **1978**, *20*, 397−405.
(51) Kellogg, G. E.; Phatak, S.; Nicholls, A.; Grant, J. A. Validation of Poisson−Boltzmann electrostatic potential fields in 3D QSAR: a CoMFA study on multiple data sets. *QSAR Comb. Sci.* **2003**, *22*, 959−964.
(52) Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri, 63144, U.S.A.
(53) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(54) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.
(55) Stone, M. Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Royal Stat. Soc. B* **1974**, *36*, 111−147.
(56) Jonathan, P.; Krzanowski, W. J.; McCarthy, W. V. On the use of cross-validation to assess performance in multivariate prediction. *Stat. Comput.* **2000**, *10*, 209−229.
(57) Hormann, R. E.; Dinan, L.; Whiting, P. Superimposition evaluation of ecdysteroid agonist chemotypes through multidimensional QSAR. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 135−153.
(58) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10−23.

CI050051L