JOC Article

# Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels−Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands

Kenny B. Lipkowitz*,§ and Meeta Pradhan†

*Department of Chemistry, Ladd Hall 104, North Dakota State University, Fargo, North Dakota 58105-5516, and Program in Chemical Informatics, Indiana University School of Informatics, 840 West Michigan Street, Indianapolis, Indiana 46202*

*kenny.lipkowitz@ndsu.nodak.edu*

*Received November 26, 2002*

A QSAR using Comparative Molecular Field Analysis (CoMFA) is developed for a set of 23 catalysts containing bisoxazoline or phosphinooxazoline ligands that are known to induce asymmetry during the Diels−Alder reaction of *N*-2-alkenoyl-1,3-oxazolidine-2-one with cyclopentadiene. It is shown that extremely high $q^2$ statistics can be derived by using standard modeling protocols when internal validation alone is done as well as when an external test set is used. From these models it is shown that approximately 70% of the variance in the observed enantiomeric excess can be attributed to the steric field and the remainder of the variance to the electrostatic field. Suggestions about how to improve the performance of inefficient catalysts are given.

## Introduction

The topic of chirality transcends traditional boundaries separating chemistry from biology and science from technology. The word "chiroscience" is sometimes used to connote the interconnectedness of science and technology as an overarching theme in studies related to understanding and using chirality.[1] This interconnectedness is especially true in the subdisciplines of the chemical sciences dedicated to the production of optically pure materials, impelled primarily by the pharmaceuticals industry that has a mandate to test enantiomeric drugs as separate entities, but also from other sectors of business.

One can obtain optically pure materials from nature or one can separate enantiomers from a mixture (racemic or otherwise). Alternatively, and preferably, one can synthesize the desired material. This can be accomplished indirectly with the use of chiral auxiliaries (that first need to be appended and then removed), or directly with chiral reagents. The latter approach coupled with catalysts is the ideal method for converting a prochiral reactant into an optically pure product, but this can be accomplished only if one of the two competing reaction channels accessible to the system is shut off. There has been a great deal of work recently focusing on this aspect of synthetic chemistry.

The emphasis of chemical research in the area of developing catalysts capable of asymmetric induction is understandable, especially from an industrial perspective where one wants high yields of highly stereoselective reactions using inexpensive reagents and where the catalysts maintain their integrity for indefinite time periods. Many advances have been made in this regard and those developments have been highlighted in numerous reviews culminating in the state-of-the-art compendium *Comprehensive Asymmetric Catalysis* published in 1999.[2] These summarizing publications indicate that we are far from achieving the industrial goals desired for chiral catalysts; while some catalysts perform well, most do not, even after extensive "tweaking". Moreover, what is learned from optimization of catalytic performance in one system usually does not transfer well to other systems; there is a clear lack of guidelines or rules for chiral catalyst design.

Understanding how chiral catalysts work and then transferring that knowledge to new and improved catalysts can be a difficult task. Traditionally chemists used hand-held mechanical models to better comprehend how asymmetry is induced but more recently molecular

§ North Dakota State University.
† Indiana University School of Informatics.

(1) (a) Rouhi, A. M. Cover story and feature article: Chiral Roundup. In *C&E News* **2002**, *June 10*, 43−57. (b) Jacoby, M. Cover story and feature article: 2D Stereoselectivity. In *C&E News* **2002**, *March 25*, 43−46. (c) Stinson, S. C. Cover story and feature article: Chiral Pharmaceuticals. In *C&E News* **2001**, *October 1*, 79−97. (d) Stinson, S. C. Chiral Drugs. In *C&E News* **2000**, *October 23*, 55−78. (e) Stinson, S. C. Cover story and feature report: Chiral Drugs: Market Growth in Single-Isomer Forms Spurs Research Advances. In *C&E News* **1995**, *October 5*, 44−74. (f) Stinson, S. C. Cover story and feature article: Chiral Drugs: Single-Isomer Products Drive Development of New Syntheses and Separation Technologies. In *C&E News* **1994**, *September 19*, 38−72. (g) Stinson, S. C. Cover story and feature article: Chiral Drugs: Wave of New Enantiomer Products Set to Flood Market. In *C&E News* **1993**, *September 27*, 38−65.

(2) *Comprehensive Asymmetric Catalysis*; Jacobsen, E. N., Pfaltz, A., Yamamoto, H., Eds.; Springer-Verlag: New York, 1999; 3-vol set, ISBN 3-540-64336-2.

10.1021/jo0267697 CCC: $25.00 
Published on Web 05/13/2003

modeling tools have been used, especially quantum-based methods. The reason for doing this is that one can *quantify* differences between diastereomeric reaction pathways by computing transition structures.[3] Using quantum theories this way is common, rigorous, and useful, but there are several problems that can diminish its utility as a predictive method. First, unless one uses approximate methods such as semiempirical quantum theories or potential energy functions that emulate transition states, one is relegated to dealing with small to medium molecules. Second, there is no guarantee that the transition structures located computationally correspond to those of the actual reaction channels being followed experimentally. Moreover, in many instances there exist a range of catalyst and substrate conformers that must be evaluated. Third, one must account for changes in spin state along the reaction path if such changes in spin take place. This has been especially problematic in the study of the Katsuki−Jacobson catalyst.[4−6] Finally, if these obstacles are cleared, it can be difficult to extract information from those transition states about the intermolecular interactions giving rise to the preference of one pathway in lieu of the other that would be of use to the bench chemist needing to design improved catalysts.

There exists another modeling technique used by scientists to make structure−activity relationships that, historically, predates modern applied quantum chemistry and which has been shown to be applicable to a wide range of problems. This method is QSAR (quantitative structure−activity relationships).[7] While QSAR has its own set of deficiencies and pitfalls, it might be a useful tool for understanding how catalysts work and to predict how to make improved catalysts. In this paper we ask the following two questions: (1) Can off-the-shelf QSAR methods be used for generating mathematical models of catalytic systems of interest to synthetic chemists that are both statistically significant and predictive? The reason for asking this is because chemists are now using combinatorial methods to provide an initial set of catalysts for screening, and based on those screening results QSAR might be able to provide guidance about what next to make (or not make). (2) How do chiral catalysts work to induce asymmetry? Clearly steric factors exist that direct incoming reagents one way or another but, for a given set of catalysts, how much of this directive influence comes from repulsive/attractive steric influences and how much originates from electrostatic effects? Furthermore, can one predict how to modulate those steric and electronic influences to make improved chiral catalysts? These questions are the impetus for the research described below.

(3) *Transition State Modeling for Catalysis*; Truhlar, D. G., Morokuma, K., Eds.; ACS Symp. Ser. 721; American Chemical Society: Washington, DC, 1999.

(4) Linde, C.; Akermark, B.; Norrby, P.-O.; Svensson, M. *J. Am. Chem. Soc.* **1999**, *121*, 5083.

(5) Abashkin, Y. G.; Collins, J. R.; Burt, S. K. *Inorg. Chem.* **2001**, *40*, 4040.

(6) El-Bahraoui, J.; Wiest, O.; Feichtinger, D.; Plattner, D. A. *Angew. Chem., Int. Ed.* **2001**, *40*, 2073.

(7) (a) *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993. (b) *QSAR: Hansch Analysis and Related Approaches*; Kubinyi, H., Ed.; VCH: Weinheim, Germany, 1994. (c) Hansch, C.; Leo, A. *Exploring QSAR Fundamentals, and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

## Background

In traditional QSAR one attempts to correlate the activities of a set of molecules (that are presumed to carry out their tasks in a similar manner) with one or more attributes of those molecules. These attributes are embedded in molecular descriptors that typically describe hydrophobic, steric, and electronic features of each molecule but a descriptor can be anything that describes some feature of a molecule including graph theory indices describing atomic connectivity. These descriptors are regressed onto the activity data to generate a mathematical model that then can be used to predict the activity of as yet unsythesized analogues.

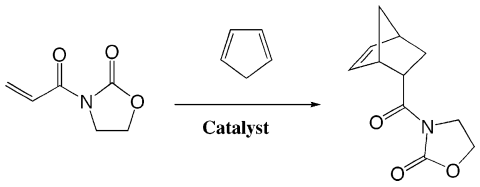
A complement to this approach is Comparative Molecular Field Analysis (CoMFA).[8] The genius of this method is that one addresses the interactions each molecule in the data set "feels" external to itself. This is done by placing each molecule, oriented the same way, at the center of a three-dimensional grid and evaluating at each grid point the interaction energy between a probe molecule (or atom) with each aligned molecule. The interaction energy at each grid point is then a descriptor that is used in the regression and, because there can be so many descriptors, robust statistical methods such as partial least squares (PLS) projections to latent variables (a method akin to principle component regression analysis) are used.[9]

In a typical QSAR or CoMFA analysis one must not only derive the mathematical model, one must then validate it. That is to say, providing a statistical measure of agreement between experiment and theory, like a correlation coefficient from a plot of computed activity (using the derived model) versus experimental activities for the molecules in the data set, is not sufficient validation. Accordingly, a variety of validation schemes have been developed, the most common of which is the leave-one-out (LOO) cross-validation method. Here one omits a single compound from the data set and a PLS model is developed by using the remaining compounds. That model is used to predict the activity of the omitted molecule that was not included in the model. This procedure is repeated until all molecules in the data set have been eliminated once. The most salient statistic from such an analysis is the cross-validated $r$ squared ($r_{cv}^2$) value, commonly published as $q^2$. A cross-validated mathematical model is important because the model is proved to be predictive. The cross-validated $r$ squared statistic, $q^2$, is always smaller than a simple $r^2$ from plots of predicted activity versus experimental activity. It is commonly stated that a statistically meaningful model has been achieved when $q^2 > 0.3-0.5$. A perusal of $q^2$ values published in the *Journal of Medicinal Chemistry* between the years 1996 and 2000 ranges from 0.56 to 0.83, with the average for 40 CoMFA analyses being $q^2 = 0.66$. We will show below that our values are substantially higher than this average.

(8) Kubinyi, H. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Editor in Chief; Wiley: New York, 1998; Vol. 1, pp 448−460, Comparative Molecular Field Analysis (CoMFA).

(9) Wold, S.; Johansson, E.; Cocchi, M. in ref 7a, pp 523−564, PLS−Partial least Squares Projections to Latent Structures.

**SCHEME 1**



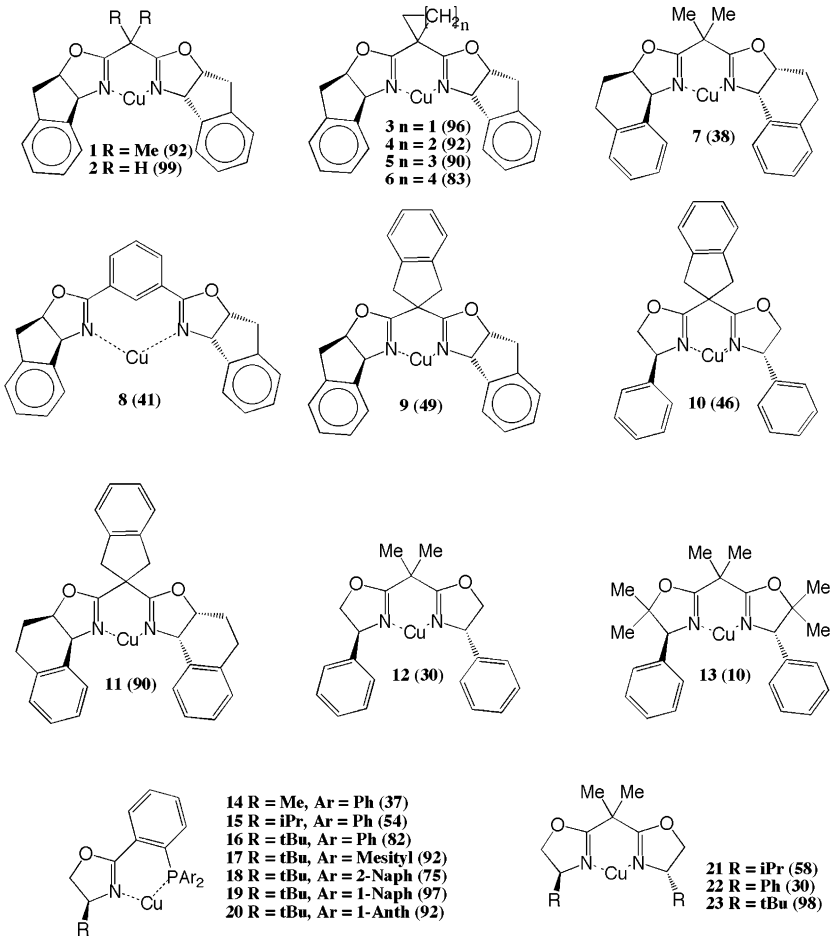

## Systems Studied

In a traditional QSAR study one might evaluate, say, the binding affinities of a series of drugs to a common receptor. The assumption being made is that all drugs bind to the same receptor and evoke biological responses via the same mechanism. In those studies the nature of the receptor is unknown. In our study we do the reverse; we know the shape of the catalysts (the receptor in our case) and we want to correlate the steric and the electrostatic fields for those chiral catalysts with the experimental enantiomeric excesses (ee) for a common reaction. The reaction we evaluate is depicted in Scheme 1. This reaction has become a de facto benchmark for synthetic chemists who want to demonstrate the utility of their chiral catalysts.

The assumption made here is that the same mechanism (a concerted pericyclic reaction of more or less similar synchonicity) prevails around the catalyst's metal center. To ensure that similar transition states exist for the reactions being studied here we select catalysts with the same metal (copper) and similar (though not redundant) ligands which have been published in the literature. Systems fulfilling the following requirements were selected for this study: (1) The publication from which we extracted our information should have a complete assessment of reaction conditions such that the reported ee is deemed reliable; the efficacy of many catalyzed reactions is often found to be dependent upon more than just the ligand used. In particular the metal (including its spin and oxidation state), the solvent, the temperature and the nature of the counterions associated with the catalyst before substrate binding all impact the observed stereochemisty. The systems we selected are all well studied in this regard by the authors who created each catalyst. The ee values used in our analyses are thus considered to be the maximum values corresponding to the optimum reaction conditions for a given ligand. (2) The reaction taking place should be as simple as possible. In the example described below we select the Diels−Alder reaction of Scheme 1 because the transition state for this concerted pericyclic reaction is not expected to change much from catalyst to catalyst.

The catalysts selected for our study are depicted in Scheme 2. Citations for each catalyst are compiled in the reference section of this paper.[10−16] All compounds selected contain at least one oxazole ring that was used for alignment (see below). It is to be noted that two basic chemotypes are contained in the table: one containing a bisoxazoline substructure and another with a phosphi-

**SCHEME 2**



nooxazoline substructure. Clearly the size, shape, and electrical properties of these molecules covers a suitable range of values for our analysis. The experimental values are in parentheses adjacent to the corresponding catalyst number. These data span a range of ee values from 10 to 99. Why and how these structurally related systems give rise to such a diverse range of ee values is not obvious and is the focus of this research.

## Computational Methods

All computations were done with commercially available software. In particular the quantum mechanical calculations were done with the PM3tm Hamiltonian implemented in Spartan 5.1.3.[17] Crystal structures (when available) were retrieved from the Cambridge Structural Database,[18] and the CoMFA was done with SYBYL 6.8.[19]

**(a) Generation of Catalyst Structures.** The geometry of each catalyst was determined quantum chemically. Initial atomic coordinates were generated with the builder facility of Spartan or they were imported from the CSD. In all cases the corresponding anions (usually triflates or antimony hexafluorides) were included in the calculations. No continuum model for solvent was applied in the calculation and the optimization was stopped when Spartan's default convergence criterion was met. In some instances more than one conformation of catalyst is possible. Following a systematic (grid) conformer search, the lowest energy conformer was used for the CoMFA.

**(b) Catalyst Alignment.** The lowest energy structure of each catalyst was imported into SYBYL (minus the associated counterions). There exist several possible alignment schemes and we opted for the most obvious: aligning the molecules by least-squares fitting the five atoms in the oxazoline ring that is common to all catalysts. Figure 1 depicts all 23 molecules that have been superimposed this way.

**(c) Selection of Variables.** Once the molecules in the data set are aligned they are centered in a three-dimensional grid with uniformly placed grid points. At each grid point a test probe is selected and the intermolecular energy between the probe and each molecule is calculated. These interaction energies are the descriptors used by SYBYL for the PLS regression. Many options are available for selection of grid dimensions, atomic charge assignments, treatment of the dielectric between probe and molecule, the probe to be used, etc. In this study we wanted to know if standard, commercially available software is amenable to CoMFA analysis of chiral catalysts. Thus we adopted a minimalist approach of (1) beginning with suggested, default settings and atomic probes in SYBYL and (2) not exploring all possible combinations of SYBYL variables in an attempt to seek the optimum CoMFA. Basically we present here a modest attempt to generate high-quality CoMFA models without exhaustively searching all possible combinations of variables, i.e., something a bench chemist is willing to do. Tabulated in the Results and Discussion section below are the lists of variables used and the corresponding validation statistics.

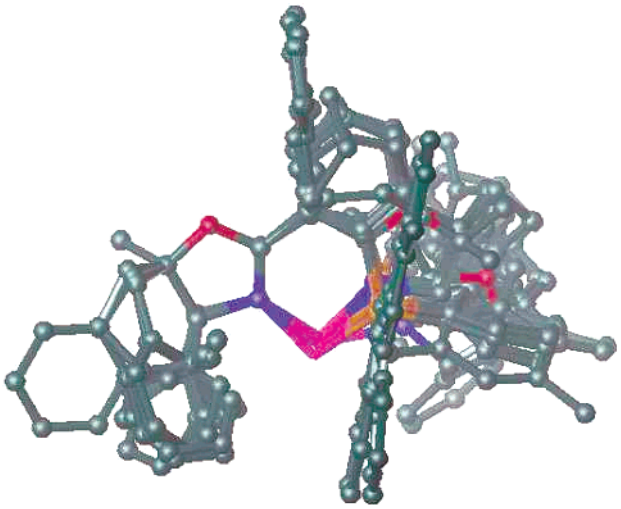


**FIGURE 1.** Alignment of all catalysts depicted in Scheme 1. Hydrogen atoms have been omitted for clarity; only the lowest energy conformer is considered in the set of catalysts.

**(d) Internal vs External Validation.** Most published CoMFA studies include some type of cross-validation, usually by invoking a leave-one-out strategy (LOO). The authors of those papers feel confident that this internal validation is satisfactory to prove the merits of the model, but Golbraikh and Tropsha[20] published a paper recently entitled "Beware of $q^2$"! In this paper the authors pointed out pitfalls of relying on internal validations alone. They demonstrated convincingly that mathematical models that are not evaluated by using an external test set (i.e., a set of compounds not included in the training set) might not be valid. They illustrated this problem by presenting several published examples of QSARs where internal validation alone gave high statistical significance but when tested against an external test set did very poorly. They argue correctly that the following three criteria must be met to adequately validate the model. First, a high correlation coefficient between predicted values and experimental values for an external test set must be obtained. Second, the slope of the plot of predicted versus actual values for that test set must be close to unity. Third, the intercept of that plot should be close to zero. These are stringent criteria for proving that a mathematical model is valid, and they are conditions that we have met (see below). Accordingly we carried out two CoMFA analyses in this paper. One divides the 23 catalysts into a training set and a test set and the other treats all 23 molecules as a single, internally validated set. For the external validation four catalysts were selected randomly (compounds **10**, **15**, **18**, and **22**) and the remaining 18 served as the training set. Note that in the external test set two compounds are bisoxazolines and two happen to be phosphinooxazolines, thus representing well the classes of catalysts.

## Results and Discussion

**(a) CoMFA with Internal Validation.** By using the aligned data set depicted in Figure 1 a standard CoMFA was performed. Several kinds of partial atomic charges were used initially for this evaluation including Mulliken and potential derived charges, but the best results were

(10) Catalysts **1**, **3**, **4**, **5**, **6**: Davies, I. W.; Gerena, L.; Cai, D.; Larsen, R. D.; Verhoeven, T. R.; Reider, P. J. *Tetrahedron Lett.* **1997**, *38*, 1145.
(11) Catalyst **2**: Ghosh, A. K.; Mathivanan, P.; Cappielo, J. *Tetrahedron Lett.* **1996**, *37*, 3815.
(12) Catalysts **7**, **9**, **10**, **11**, **12**: Davies, I. W.; Gerena, L.; Cai, D.; Larsen, R. D.; Verhoeven, T. R.; Reider, P. J. *Tetrahedron Lett.* **1997**, *38*, 1145.
(13) Catalyst **8**: Davies, I. W.; Senanayake, C. H.; Larsen, R. D.; Verhoeven, T. R.; Reider, P. J. *Tetrahedron Lett.* **1996**, 1725.
(14) Catalyst **13**: Evans, D. A.; Miller, S. J.; Lectka, T.; von Matt, P. *J. Am. Chem. Soc.* **1999**, *121*, 7559.
(15) Catalysts **14**, **15**, **16**, **17**, **18**, **19**, **20**: Sagasser, I.; Helmchen, G. *Tetrahedron Lett.* **1998**, *39*, 261.
(16) Catalysts **21**, **22**, **23**: Evans, D. A.; Lectka, T.; Miller, S. J. *Tetrahedron Lett.* **1993**, *34*, 7027.
(17) Wavefunction, Inc.: 18401 Von Karman Ave., Suite 370, Irvine, CA 92715.
(18) Cambridge Structural Database. Available from Wavefunction, Inc.: 18401 Von Karman, Suite 370, Irvine, CA 92715.
(19) Tripos Associates Inc.: 1699 South Hanley Road, St. Louis, MO 63144 USA.
(20) Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.

**TABLE 1. Influence of Selected Variables on CoMFA Results for 23 Catalysts**

| fields[a] | energy cutoff[b] | dielectric function | grid spacing[c] | probe type | latent variables[d] | $r_{cv}^2$ | $r^2$ [e] |
|---|---|---|---|---|---|---|---|
| | | | field | | | | |
| both | 30/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.804//0.707 | 0.991 |
| both | 30/20 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.833//0.732 | 0.977 |
| both | 30/10 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.836//0.734 | 0.992 |
| both | 30/5 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.832//0.729 | 0.993 |
| E | 30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.688//0.626 | 0.994 |
| S | 30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.784//0.659 | 0.995 |
| both | 20/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.779//0.680 | 0.995 |
| both | 10/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.774//0.656 | 0.994 |
| both | 5/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.776//0.626 | 0.988 |
| | | | column filtering[f] | | | | |
| both | 30/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.813//0.721 | 0.991 |
| both | 30/20 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.833//0.746 | 0.994 |
| both | 30/10 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.810//0.720 | 0.993 |
| both | 30/5 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.832//0.729 | 0.993 |
| both | 20/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.793//0.697 | 0.995 |
| both | 10/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.789//0.683 | 0.994 |
| both | 5/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.829//0.775 | 0.988 |
| | | | dielectric | | | | |
| both | 30/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.825//0.717 | 0.994 |
| both | 30/20 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.826//0.715 | 0.994 |
| both | 30/10 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.822//0.721 | 0.993 |
| both | 30/5 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.808//0.717 | 0.993 |
| both | 20/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.808//0.717 | 0.993 |
| both | 10/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.786//0.682 | 0.992 |
| both | 5/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.786//0.659 | 0.990 |
| both[f] | 30/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.838//0.720 | 0.994 |
| both[f] | 30/20 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.801//0.702 | 0.994 |
| both[f] | 30/10 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.755//0.661 | 0.993 |
| both[f] | 30/5 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.810//0.686 | 0.993 |
| both[f] | 20/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.820//0.717 | 0.993 |
| both[f] | 10/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.814//0.696 | 0.992 |
| both[f] | 5/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.840//0.772 | 0.990 |
| | | | grid spacing | | | | |
| both | 30/20 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.833//0.732 | 0.977 |
| both | 30/20 | $1/r^2$ | 1.75 | $C^+_{sp^3}$ | 6//3 | 0.717//0.598 | 0.990 |
| both | 30/20 | $1/r^2$ | 1.50 | $C^+_{sp^3}$ | 6//3 | 0.729//0.632 | 0.992 |
| both | 30/20 | $1/r^2$ | 1.0 | $C^+_{sp^3}$ | 6//3 | 0.696//0.585 | 0.994 |
| both | 30/20 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.630//0.510 | 0.992 |
| both | 30/20 | $1/r$ | 1.75 | $C^+_{sp^3}$ | 6//3 | 0.697//0.563 | 0.994 |
| both | 30/20 | $1/r$ | 1.50 | $C^+_{sp^3}$ | 6//3 | 0.734//0.606 | 0.993 |
| both | 30/20 | $1/r$ | 1.00 | $C^+_{sp^3}$ | 6//3 | 0.692//0.565 | 0.994 |
| | | | probe type | | | | |
| both | 30/20 | $1/r^2$ | 2.00 | $H^+$ | 6//3 | 0.702//0.604 | 0.987 |
| both | 30/20 | $1/r^2$ | 1.75 | $H^+$ | 6//3 | 0.717//0.598 | 0.990 |
| both | 30/20 | $1/r^2$ | 1.50 | $H^+$ | 6//3 | 0.739//0.644 | 0.993 |
| both | 30/20 | $1/r^2$ | 1.00 | $H^+$ | 6//3 | 0.726//0.629 | 0.993 |
| both | 30/20 | $1/r$ | 2.00 | $H^+$ | 6//3 | 0.671//0.557 | 0.987 |
| both | 30/20 | $1/r$ | 1.75 | $H^+$ | 6//3 | 0.686//0.567 | 0.991 |
| both | 30/20 | $1/r$ | 1.50 | $H^+$ | 6//3 | 0.714//0.620 | 0.994 |
| both | 30/20 | $1/r$ | 1.00 | $H^+$ | 6//3 | 0.700//0.599 | 0.993 |
| both | 30/20 | $1/r^2$ | 2.00 | $O^-_{sp^3}$ | 6//3 | 0.684//0.610 | 0.989 |
| both | 30/20 | $1/r^2$ | 1.75 | $O^-_{sp^3}$ | 6//3 | 0.697//0.602 | 0.992 |
| both | 30/20 | $1/r^2$ | 1.50 | $O^-_{sp^3}$ | 6//3 | 0.756//0.653 | 0.993 |
| both | 30/20 | $1/r^2$ | 1.00 | $O^-_{sp^3}$ | 6//3 | 0.739//0.638 | 0.995 |
| both | 30/20 | $1/r$ | 2.00 | $O^-_{sp^3}$ | 6//3 | 0.619//0.533 | 0.989 |
| both | 30/20 | $1/r$ | 1.75 | $O^-_{sp^3}$ | 6//3 | 0.684//0.563 | 0.992 |
| both | 30/20 | $1/r$ | 1.50 | $O^-_{sp^3}$ | 6//3 | 0.731//0.625 | 0.993 |
| both | 30/20 | $1/r$ | 1.00 | $O^-_{sp^3}$ | 6//3 | 0.704//0.595 | 0.990 |

[a] Fields are steric (S) and electrostatic (E) or both. [b] Steric cutoff listed first/Coulomb cutoffs listed second, values in kcal/mol. [c] Angstrom. [d] optimum/minimum number of components. [e] These values are for three latent variables only. [f] Column filtering set to 2.

generally obtained with Gasteiger charges. The results described herein are thus based on Gasteiger charges. Table 1 summarizes some of our results. The headings in each column refer to the following: the type of field used for the analyses, how the dielectric of the medium was treated, the spacing of the uniform grid in which the aligned molecules are embedded, the type of probe used at each grid point to calculate the interaction energies, the number of latent variables extracted by the PLS projection, the magnitude of the cross-validated *r* squared coefficient, and finally, the simple correlation coefficient, r squared, for that model.

The first entry in this table is the field used in the analysis. The fields used in this study are the steric field and the electrostatic field. It is to be noted that other kinds of fields could be included, e.g., hydrogen bonding,

hydrophobic, etc., but these two fields represent best the interactions that take place between this set of chiral catalysts and the reagents undergoing the Diels−Alder reaction so they are used exclusively. When both fields are included in the analysis there are twice as many descriptors as when electrical fields or steric fields are used alone. The next column in the table is the magnitude of the energy cutoff used. The energy cutoffs need some explanation—at each grid point the interaction energy between a probe atom and each of the atoms in each of the aligned molecules is computed with a potential energy function. In some instances the grid points lie close to one or more atoms and consequently the computed energies at those points are unacceptably large. Thence we invoke a cutoff that says, for example, remove all descriptors that exceed a repulsive energy of 30 kcal/mol for sterics, and likewise remove data points containing Coulombic attraction or repulsion of, say, 20 kcal/mol. Various combinations of cutoffs are listed in the table to illustrate the sensitivity of the model to such omissions. The dielectric treatment we used initially is $1/r^2$ and the grid spacing was set to 2.0 Å. The initial probe was a tetrahedral "carbon atom" with a +1 charge for purposes of computing the Lennard-Jones steric field and the Coulomb electric fields, respectively. In the column labeled "latent variables" (LVs) we provide the optimum number of variables determined from the PLS analysis followed by (//) the minimum number of latent variables. In traditional multiple regression methodology one uses ~5 observations per term in the QSAR equation. The reason for this is that the molecular descriptors used for QSARs are not orthogonal to each other. Hence, using those nonorthogonal descriptors one would require about 4 terms in the QSAR model for the 23-molecule data set. Contrarily, because the latent variables of a PLS are orthogonal, one can use as many LVs as needed to provide the best model. Nonetheless, to be conservative we provide information about models using the minimum number of LVs in addition to the optimum number. The penultimate column in the table lists the $r_{cv}^2$ from LOO cross-validation and the last column is the squared correlation coefficient, $r^2$. The first number in this column corresponds to the optimum number of latent variables followed by (//) the value for only 3 latent variables.

The first entries in Table 1, listed under "field", are the result of models constructed where the cutoff values of the fields have been changed. In this subsection of Table 1 we find the best model to be that where the steric cutoff is at 30 kcal/mol and the electric cutoff is at 10 kcal/mol. A $q^2$ value of 0.836 for 6 latent variables (0.734 for 3 LVs) is an extremely high number when compared to the recent literature. Indeed, this value for a CoMFA far exceeds the average published in the *Journal of Medicinal Chemistry* as described earlier and is comparable to one of the highest values presented in that journal. Regression is an ill-posed problem in statistics that can result in models that are not sufficiently stable when trained on noisy data. Because many grid points may have equivalent energies (for example, sampling interaction energies at points at the extremities of the grid far from the aligned molecules would give rise to small Lennard-Jones interaction energies of comparable values) one is introducing a significant amount of "noise" rather than "signal". We remove points with only minor variation in their field values using SYBYL's minimum-$\sigma$ condition. This deletes points having a lower variance than that assigned (in this work, 2). Removing points with only minor variation in the field improved some of the models.

The next variable we considered was how best to treat the dielectric for the electrostatic interaction. Earlier entries assumed a distance dependence of $1/r^2$. In the subsection of Table 1 entitled "dielectric" we considered a $1/r$ dependency. Although some changes in $r_{cv}^2$ are noted, these changes are not substantive suggesting that the electrostatic contributions to these models are relatively minor and that most of the contribution to stereoinduction originates from steric interactions. It will be shown below that this foreshadowing is, indeed, correct.

The next variable we considered is the number of descriptors generated. As the grid spacing becomes smaller the number of data points increases. Too many descriptors can actually degrade a PLS performance so we changed the grid spacing only within the range 1−2 Å. What we find in Table 1 is that the models become less significant as the number of descriptors increases when a dielectric treatment of $1/r^2$ is used. Contrarily there is an improvement in results when the grid coarseness is reduced for a $1/r$ dielectric. In either event, the results are not as good as those described above with spacing of 2 Å. Finally, we considered alternative probe atoms for generating descriptors as the last set of entries in Table 1. We consider two alternative probes with different treatments of dielectric and for different grid spacing. No improvements are noted. At this point, then, we have several models that are internally predictive with very high values of $q^2$ that we shall come back to later. First, however, we need to consider what would happen if the CoMFA were constructed and then tested with an external test set.

**(b) CoMFA with External Validation.** To perform this assessment we selected, randomly, 4 of the 23 compounds to serve as an external test set (catalysts **10**, **15**, **18**, and **22**) and used the remaining 18 molecules to perform the CoMFA analysis. The same strategy described above was used here. The results are compiled in Table 2 where the column headings and the column subsections have the same meaning as described for Table 1.

Because fewer compounds are being used in this CoMFA one anticipates and one sees a degradation of the quality of the models derived. Still, the results are extremely good with $r_{cv}^2$ values approaching 0.8 for five LVs and 0.7 for three LVs. Again we emphasize that these are predictive models but they are validated internally only. Using the best model we then predicted the ee values of the four compounds omitted from the training set. A plot of computed ee versus experimental ee gave a linear relationship with a squared correlation coefficient, $r^2$, of 0.94. Most importantly, in concordance with the criteria highlighted by Golbraikh and Tropsha, we find a slope near unity (1.03) and an intercept of 7.5. While the latter is somewhat elevated we find that the model derived from the test set satisfies well Tropsha's arguments for a statistically valid QSAR.

Given a relatively small set of catalysts (18), we are thus able to predict the ee of new catalysts. Hence we have shown here that commercially available software

**TABLE 2. Influence of Selected Variables on CoMFA Results for a Training Set of 19 Catalysts**

| fields[a] | energy cutoff[b] | dielectric function | grid spacing[c] | probe type | latent variables[d] | $r_{cv}^2$ | $r^2$ [e] |
|---|---|---|---|---|---|---|---|
| | | | field | | | | |
| both | 30/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.763//0.635 | 0.998 |
| both | 30/20 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.785//0.653 | 0.998 |
| both | 30/10 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.784//0.672 | 0.998 |
| both | 30/5 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.779//0.673 | 0.999 |
| E | 30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.666//0.573 | 0.997 |
| S | 30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.745//0.663 | 0.998 |
| both | 20/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.739//0.632 | 0.999 |
| both | 10/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.731//0.621 | 0.999 |
| both | 5/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.723//0.566 | 0.999 |
| | | | column filtering[f] | | | | |
| both | 30/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.764//0.650 | 0.998 |
| both | 30/20 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.764//0.662 | 0.998 |
| both | 30/10 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.734//0.648 | 0.998 |
| both | 30/5 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.713//0.615 | 0.999 |
| both | 20/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.758//0.663 | 0.999 |
| both | 10/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.731//0.621 | 0.999 |
| both | 5/30 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 3 | 0.615 | 0.999 |
| | | | dielectric | | | | |
| both | 30/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.773//0.671 | 0.994 |
| both | 30/20 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.789//0.682 | 0.999 |
| both | 30/10 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.792//0.692 | 0.997 |
| both | 30/5 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.781//0.689 | 0.999 |
| both | 20/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.784//0.681 | 0.999 |
| both | 10/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.756//0.653 | 0.999 |
| both | 5/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.742//0.621 | 0.999 |
| both[f] | 30/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.783//0.678 | 0.999 |
| both[f] | 30/20 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.695//0.647 | 0.999 |
| both[f] | 30/10 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 4//3 | 0.608//0.555 | 0.999 |
| both[f] | 30/5 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.668//0.544 | 0.999 |
| both[f] | 20/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.811//0.702 | 0.999 |
| both[f] | 10/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.779//0.690 | 0.999 |
| both[f] | 5/30 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 4//3 | 0.699//0.671 | 0.999 |
| | | | grid spacing | | | | |
| both | 30/20 | $1/r^2$ | 2.00 | $C^+_{sp^3}$ | 6//3 | 0.785//0.653 | 0.998 |
| both | 30/20 | $1/r^2$ | 1.75 | $C^+_{sp^3}$ | 6//3 | 0.618//0.439 | 0.998 |
| both | 30/20 | $1/r^2$ | 1.50 | $C^+_{sp^3}$ | 6//3 | 0.671//0.521 | 0.998 |
| both | 30/20 | $1/r^2$ | 1.0 | $C^+_{sp^3}$ | 6//3 | 0.605//0.465 | 0.998 |
| both | 30/20 | $1/r$ | 2.00 | $C^+_{sp^3}$ | 5//3 | 0.789/0.682 | 0.999 |
| both | 30/20 | $1/r$ | 1.75 | $C^+_{sp^3}$ | 5//3 | 0.561//0.423 | 0.999 |
| both | 30/20 | $1/r$ | 1.50 | $C^+_{sp^3}$ | 5//3 | 0.660//0.530 | 0.999 |
| both | 30/20 | $1/r$ | 1.00 | $C^+_{sp^3}$ | 5//3 | 0.632//0.424 | 0.999 |
| | | | probe type | | | | |
| both | 30/20 | $1/r^2$ | 2.00 | $H^+$ | 5//3 | 0.620//0.505 | 0.997 |
| both | 30/20 | $1/r^2$ | 1.75 | $H^+$ | 6//3 | 0.663//0.500 | 0.999 |
| both | 30/20 | $1/r^2$ | 1.50 | $H^+$ | 5//3 | 0.682//0.569 | 0.999 |
| both | 30/20 | $1/r^2$ | 1.00 | $H^+$ | 5//3 | 0.658//0.531 | 0.999 |
| both | 30/20 | $1/r$ | 2.00 | $H^+$ | 5//3 | 0.615//0.517 | 0.998 |
| both | 30/20 | $1/r$ | 1.75 | $H^+$ | 6//3 | 0.655//0.522 | 0.999 |
| both | 30/20 | $1/r$ | 1.50 | $H^+$ | 5//3 | 0.691//0.584 | 0.999 |
| both | 30/20 | $1/r$ | 1.00 | $H^+$ | 5//3 | 0.646//0.534 | 0.996 |
| both | 30/20 | $1/r^2$ | 2.00 | $O^-_{sp^3}$ | 5//3 | 0.608//0.512 | 0.999 |
| both | 30/20 | $1/r^2$ | 1.75 | $O^-_{sp^3}$ | 6//3 | 0.603//0.468 | 0.998 |
| both | 30/20 | $1/r^2$ | 1.50 | $O^-_{sp^3}$ | 5//3 | 0.680//0.573 | 0.998 |
| both | 30/20 | $1/r^2$ | 1.00 | $O^-_{sp^3}$ | 5//3 | 0.658//0.521 | 0.997 |
| both | 30/20 | $1/r$ | 2.00 | $O^-_{sp^3}$ | 5//3 | 0.589//0.506 | 0.999 |
| both | 30/20 | $1/r$ | 1.75 | $O^-_{sp^3}$ | 6//3 | 0.614//0.485 | 0.999 |
| both | 30/20 | $1/r$ | 1.50 | $O^-_{sp^3}$ | 5//3 | 0.680//0.565 | 0.999 |
| both | 30/20 | $1/r$ | 1.00 | $O^-_{sp^3}$ | 5//3 | 0.646//0.523 | 0.999 |

[a] Fields are Steric (S) and Electrostatic (E) or Both. [b] Steric cutoff listed first/Coulomb cutoffs listed second, values in kcal/mol. [c] Angstrom. [d] Optimum/Minimum number of components. [e] These values are for three latent variables only. [f] Column filtering set to 2.

can be readily implemented for this purpose. This is important because this size data set and this range of ee values is about what one would find from an initial combi-chem evaluation of chiral catalysts. From these limited data one can then make some predictions about how good or how poor a given catalyst will be that has yet to be made.

The second part of this research project is meant to help explain how these catalysts work. In particular we want to know, in a quantitative fashion, what role steric and electrostatic factors play in asymmetric induction. Inspection of mechanical models provides no quantifiable information about this. Computational models are better but they too suffer from being difficult to derive quantitative information. For example, in Figure 2 are plots of

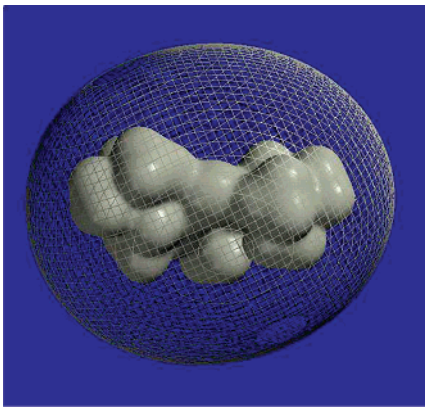


**FIGURE 2.** Electrostatic potential (grid) surrounding the van der Waals surface of catalyst **23**.

van derWaals and electrostatic surfaces of the $Cu^{2+}$ bisoxazoline, **23**.

Although the electrostatic potential surrounding the catalyst in Figure 2 looks symmetric, it is not (like the van der Waals surface it too is chiral), and we want to know what percent of the asymmetric induction for such catalysts can be attributed to these subtle, chiral electrical effects. The van der Waals surface has better defined grooves and cavities in which stereoinduction can arise but, as with the electrostatic plot, quantifying this is difficult. What is needed is information concerning steric and electrostatic modes of stereoinduction. Specifically we want to know how much of the stereoinduction arises from steric effects and how much comes from electrical effects.

CoMFA can provide insights concerning these issues. Moreover, one can also use the CoMFA model as a guide for synthesis. In this study, as in any QSAR, we want to find a linear relationship that relates an activity (in this case the ability of a catalyst to induce asymmetry during a Diels−Alder reaction) to the intensity of the surrounding fields. From our best models we find that 60−70% of the variance in the data can be described by the steric field. The remaining 30−40% of the variance is attributed to the electrostatic field. This can be loosely interpreted as meaning that most of the stereoinduction originates from steric effects of the ligands surrounding the catalyst. While this seems intuitive, and while it is fully consonant with our perceptions of how these particular catalysts work, we point out that we are able to provide a *quantitative* assessment of how much each effect contributes to the stereoinduction. An intuitive evaluation for other catalysts may not be so obvious, however, when electrical effects are as important as (or more so than) steric effects. Hence models such as ours have the potential for assisting in the construction of improved catalysts where both steric bulk and electronics can be modulated via synthesis.

Another advantage of using a QSAR like that presented here is that one can visualize the large number of computed descriptor coefficients by making iso-value contour maps of those coefficients at grid points surrounding the aligned data set. Rather than using the coefficients themselves we present a more common "standard deviation times coefficient" plot (STDV*COEFFICIENTS), where, at each grid point, the

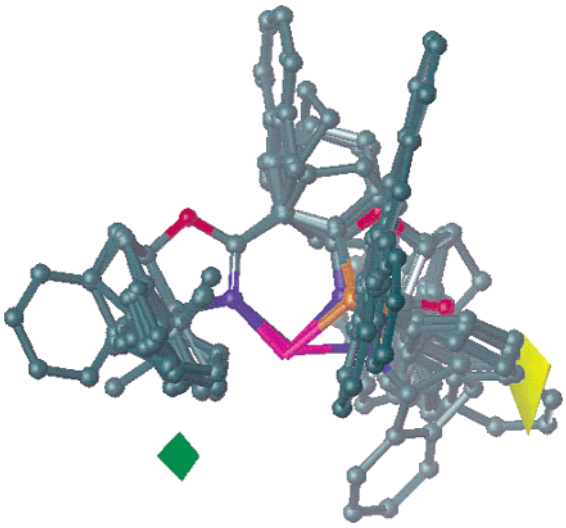


**FIGURE 3.** CoMFA steric STDEV*COEFF contour plot. Shown inside the field is the aligned set of 23 chiral catalysts with hydrogen atoms removed for clarity. Placement of bulky groups near the green region (contoured at contribution level 93) and/or removal of steric bulk near the yellow region (contoured at contribution level 7) should increase ee for those catalysts that are not very stereoselective.

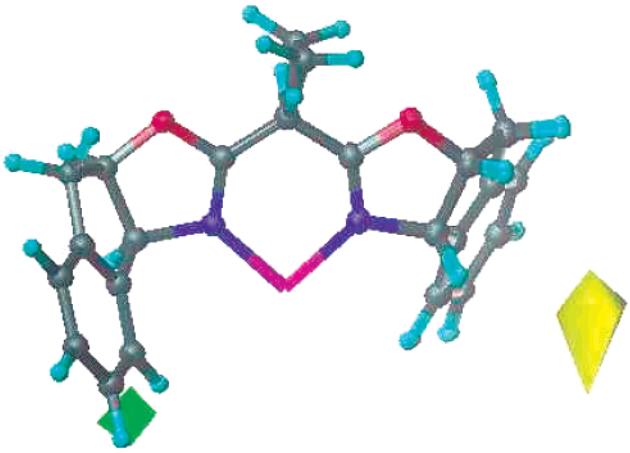


**FIGURE 4.** CoMFA steric STDEV*COEFF contour plot. Shown inside the field is the highly efficient catalyst **3** (ee 96%). It is to be noted that significant steric bulk lies in the green region while the yellow region is devoid of steric bulk confirming the model.

standard deviation of the energies for all compounds is multiplied by the PLS coefficient. Plotted in Figure 3 are regions of space where steric bulk should enhance or destroy stereoinduction. We focus here on steric influences because most of the variance is explained with the steric field. In Figure 3 is the aligned set of 23 catalysts.

To increase enantioselectivity the models indicate that more bulk is to be placed in the space encapsulated in green and steric bulk is to be removed from the region encapsulated in yellow. If these plots are meaningful one should find that highly efficient catalysts already have steric bulk in the green region and are already devoid of bulk in the yellow region. In Figure 4 we present one such efficient catalyst, **3**, where this is clearly seen.

Furthermore, one would expect that inefficient chiral catalysts either lack steric bulk in the green region and/

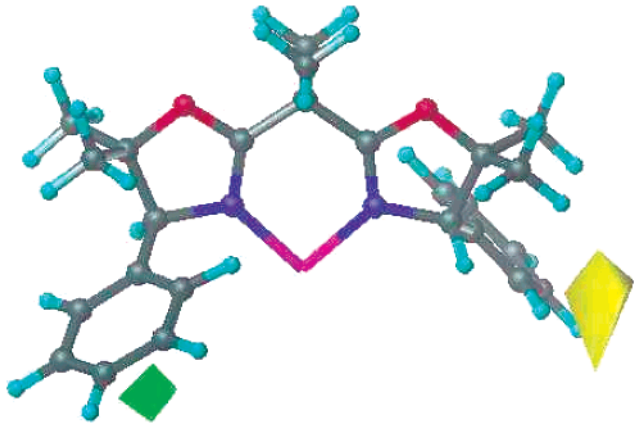


**FIGURE 5.** CoMFA steric STDEV*COEFF contour plot. Shown inside the field is the inefficient catalyst **13** (ee 10%). It is to be noted that while significant steric bulk lies in the green region the yellow region has too much steric bulk that, in turn, reduces the effectiveness of this catalyst.

or have too much steric bulk in the yellow region. In Figure 5 we show catalyst **13** (ee 10). In this figure it appears that the phenyl group on the left-hand side of the diagram is too small to be effective while the phenyl group on the right is too large. Modulating bulk such that more steric pressure exists on the left-hand side, and less steric pressure exists on the right-hand side is being suggested by our model as a means of enhancing stereoselectivity.

Similar arguments concerning the influence of electrostatic interactions can be given, but to conserve space, we do not present these plots here (in all models the electrical interactions are much less important than are the steric interactions). Finally, while we are illustrating these regions for a single model, we point out that similar (although not identical) plots for the other high $q^2$ models give qualitatively similar results.

## Summary

There were two goals set forth at the beginning of our work. The first was to see if one could use off-the-shelf software to generate high-quality QSARs for a small data set of chiral catalysts. It was shown that extremely high cross-validated $r^2$ values could be generated quickly, and with undue difficulty. The importance of this is that the set of compounds used here is comparable to a typical small combinatorial library that one might develop in an exploratory research endeavor. We have shown that such small libraries are amenable to such modeling. In this work we carried out two CoMFAs. One included all 23 catalysts with internal validation only. The second study divided the catalysts into a training set that was used to develop the mathematical model and a test set that was used to validate the model. Moreover, we demonstrated here that the external test set gave results for which a plot of predicted ee values versus observed ee values had a high coefficient of regression, a slope near unity, and an intercept that is reasonably close to zero, thus fulfilling the stringent requirements for a statistically valid mathematical model set forth by Golbraikh and Tropsha.

The second goal of the work was to better understand, qualitatively and quantitatively, why some catalysts work efficiently at asymmetric induction for the reaction in Scheme 1 while others do not. Quantitatively we are able to predict with a high degree of accuracy which catalysts are effective at carrying out this stereinduction and which are not. Quantitatively we were also able to show that approximately 70% of the variance in the model arises from the steric field while the remaining 30% is electrostatic in nature. Quantitatively, we were able to define regions in space where steric bulk will influence the outcome of the reaction. Using this model, then, we could now compute the shapes of other copper-coordinated ligands and predict their activity, i.e., we are in the position of doing statistically meaningful computer-aided molecular design (CAMD).

**Acknowledgment.** This work was carried out by grants to Indiana University from the National Science Foundation (CHE-9982888) and the Petroleum Research Fund (35172-AC4), administered by the American Chemical Society. This work was carried out on the IUPUI campus.

JO0267697