Inorg. Chem. **2007**, *46*, 8613–8624

Use of Drug Discovery Tools in Rational Organometallic Catalyst Design

Michael L. Drummond*,† and Bobby G. Sumpter‡

Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831

Received April 7, 2007

A computational procedure is detailed where techniques common in the drug discovery process—2D- and 3D-quantitative structure–activity relationships (QSAR)—are applied to rationalize the catalytic activity of a synthetically flexible, Ti–N═P ethylene polymerization catalyst system. Once models relating molecular properties to catalyst activity are built with the two QSAR approaches, two database mining approaches are used to select a small number of ligands from a larger database that are likely to produce catalysts with high activity when grafted onto the Ti–N═P framework. The software employed throughout this work is freely available, is easy to use, and was applied in a "black box" approach to highlight areas where the drug discovery tools, designed to address organic molecules, have difficulty in addressing issues arising from the presence of a metal atom. In general, 3D-QSAR offers an efficient way to screen new potential ligands and separate those likely to lead to poor catalysts from those that are likely to contribute to highly active catalysts. The results for 2D-QSAR appear to be quantitatively unreliable, likely due to the presence of a metal atom; nonetheless, there is evidence that qualitative predictions from different models may be reliable. Pitfalls in the database mining techniques are identified, none of which are insurmountable. The lessons learned about the potential uses and drawbacks of the techniques described herein are readily applicable to other catalyst frameworks, thereby enabling a rational approach to catalyst improvement and design.

I. Introduction

A recent analysis[1] of the Cambridge Structural Database (CSD) found that 70 of the most common ligands in organometallic crystal structures account for roughly 70% of the total ligands reported. Conversely, 19 000 out of a total of 22 000 distinct ligands were reported less than 10 times, indicating that, in synthesizing organometallic complexes, chemists favor the same group of ligands. While it is unsurprising that the chemist's synthetic toolkit consists of well-known, well-behaved, and well-characterized ligands leading to well-understood catalysts, a failure to utilize ligands outside of this toolkit is extremely limiting. The use of novel ligands in catalytic systems may open up new possibilities, for example, for improved activities, selectivities, solubilities, product polymer tacticity, range of reactants, etc.

Perhaps one of the most striking examples of the benefits obtained through the use of "uncommon" ligands is found in the replacement of the ubiquitous phosphine ligand with the N-heterocyclic carbene (NHC) framework.[2−8] Following the initial synthesis of isolable carbenes,[9] similarities were postulated between the chemistry of more traditional electron-rich phosphines and the newer NHCs.[10−12] Spectroscopic studies[13−14] confirmed the similarities, as both ligand groups are strong σ-donors and weak π-acceptors and, therefore, could be substituted for one another. Today, so-called "second-generation Grubbs catalysts", where NHCs have

* To whom correspondence should be addressed. E-mail: drummondml@gmail.com.

† Computer Science and Mathematics Division, Oak Ridge National Laboratory.

‡ Computer Science and Mathematics Division and Center for Nanophase Materials Science, Oak Ridge National Laboratory.

(1) Harris, S. E.; Orpen, A. G.; Bruno, I. J.; Taylor, R. *J. Chem. Inf. Model.* **2005**, *45*, 1727.

(2) Nguyen, S. T.; Johnson, L. K.; Grubbs, R. H.; Ziller, J. W. *J. Am. Chem. Soc.* **1992**, *114*, 3974.

(3) Nguyen, S. T.; Grubbs, R. H.; Ziller, J. W. *J. Am. Chem. Soc.* **1993**, *115*, 9858.

(4) Schwab, P.; Grubbs, R.; Ziller, J. W. *J. Am. Chem. Soc.* **1996**, *118*, 100.

(5) Fürstner, A.; Ackermann, L.; Gabor, B.; Goddard, R.; Lehmann, C. W.; Mynott, R.; Stelzer, F.; Thiel, O. R. *Chem. Eur. J.* **2001**, *7*, 3236.

(6) Herrmann, W. A. *Angew. Chem. Int. Ed.* **2002**, *41*, 1290.

(7) César, V.; Bellemin-Laponnaz, S.; Gade, L. H. *Chem. Soc. Rev.* **2004**, *33*, 619.

(8) Peris, E.; Crabtree, R. H. *Coord. Chem. Rev.* **2004**, *248*, 2239.

(9) Arduengo, A. J.; Harlow, R. L.; Kline, M. *J. Am. Chem. Soc.* **1991**, *113*, 361.

10.1021/ic700670s CCC: $37.00
Published on Web 09/14/2007

taken the place of phosphine, are recognized as generally possessing both improved reactivities and stabilities compared to the original phosphine-based ruthenium catalysts.[5] Needless to say, these benefits would not have been realized had chemists contented themselves with phosphine-based catalysts.

Similar success stories can certainly be found among the 19 000 underutilized organometallic ligands,[1] not to mention the possible useful ligands waiting to be discovered among the essentially infinite variety of ligands not yet utilized at all. However, even if the search is restricted to the smaller set of those ligands already studied and therefore found in the CSD, there still remains the task of separating the few potentially beneficial ligands from the vast majority of ligands that are of little catalytic use. The synthesis and experimental evaluation of tens of thousands of catalysts with possibly useful ligands is not practical, and while a combinatorial technique may improve the utility of such an experimental approach, this technique is generally of most use in the area of heterogeneous catalysts.[15] Therefore, the goal of the current work is to evaluate possible computational modeling procedures that are fast, low-cost, and accurate in identifying uncommon or wholly novel ligands likely to be of use in highly active catalysts.

Standing in the way of this goal are a number of challenges,[16] including possible inverse relationships between activity and selectivity; the difficulty in identifying simple rules of use to synthetic organometallic chemists; and effects besides the structure of the catalyst, such as temperature, catalyst loading, or the necessity of a cocatalyst. However, despite these challenges, the promise of a computational determination of catalytic ligand utility is alluring, and some initial forays have been made into applying computational techniques to rationalize the activity of organometallic catalysts. The most common approach, 3D-QSAR (quantitative structure−activity relationship), appears in, to date, only three published reports,[17−19] demonstrating the youth of this line of work. 3D-QSAR is routinely employed in the computer-aided drug discovery field, where the "3D" indicates that properties relatable to activity, such as steric bulk and atomic charges, are calculated from the 3D geometries of the species of interest. Using the comparative molecular field analysis (CoMFA) variety of 3D-QSAR,[17] the enantiomeric excess (rather than activity) of chiral, copper-containing Diels−Alder catalysts were correlated primarily to the steric properties of the catalysts, with lesser contributions noted from the electrostatic properties. CoMFA was also used[18,19] to explain the ethylene polymerization behavior of (primarily) zirconocene catalysts, based on not only sterics and electrostatics but also on the LUMO and local softness fields as predicted by density functional theory (DFT). Properties calculated semiempirically were correlated[20] to activities of iron−tetraphenylporphyrin complexes using a 2D-QSAR approach, where "2D" indicates that the connectivities of the complexes, rather than explicit structures in 3D space, are used to rationalize activities. A 2D-QSAR approach, again using DFT-calculated properties, was also used to investigate ruthenium metathesis catalysts.[21] This last study is particularly noteworthy as not only were the independent properties (i.e., the chemical descriptors calculated from the catalyst connectivity) calculated with DFT, but so too were the dependent variables (i.e., the catalyst activity or an approximation thereof).

In the studies just mentioned, computational methods were used to identify trends in ligand characteristics leading to beneficial catalytic properties. These trends generally take the form of, for example, "Add steric bulk at this point in space". Thus, any newly proposed ligands can be evaluated using such guidelines, leading to an efficient method for screening ligands. However, none of these previous studies provide a means for identifying *which* ligands should be screened, and thus, the problem of separating the small number of potentially useful ligands from the huge number of useless ones remains. In the current work, therefore, we propose coupling both 2D- and 3D-QSAR approaches, which rationalize ligand characteristics and provide patterns contributing to highly active catalysts, with another technique routinely employed in the drug discovery process: database mining.

Although database mining is not entirely unknown in the organometallic arena—it has been previously used, for example, to identify statistically common, and therefore ostensibly favorable, intermolecular interactions in crystal structures[22]—this study is the first where database mining is used to search for ligands that may prove useful in catalytic species. The approach taken is to mine databases for structures "similar" to ligands found in known, highly active catalysts. Similarity is defined in this work in two ways. In the first, a simple comparison of connectivity is made, thus identifying, for example, the two pentacycles imidazole and cyclopentadiene, or Cp and MeCp, as similar. In the second approach, similarity is based on general molecular properties, such as surface area or number of hydrophilic centers, and therefore, two ligands with entirely different structures can be considered similar. This latter, more general measure of similarity has been used recently[23] to categorize ligands based on DFT-calculated quantum mechanical descriptors, although

(10) Herrmann, W. A.; Mihalios, D.; Öfele, K.; Kiprof, P.; Belmedjahed, F. *Chem. Ber.* **1992**, *125*, 1795.
(11) Öfele, K.; Herrmann, W. A.; Mihalios, D.; Elison, M.; Herdtweck, E.; Scherer, W.; Mink, J. *J. Organomet. Chem.* **1993**, *459*, 177.
(12) Herrmann, W. A.; Öfele, K.; Elison, M.; Kühn, F. E.; Roesky, P. W. *J. Organomet. Chem.* **1994**, *480*, C7.
(13) Herrmann, W. A.; Runte, O.; Artus, G. R. J. *J. Organomet. Chem.* **1995**, *501*, C9.
(14) Öfele, K.; Herrmann, W. A.; Mihalios, D.; Elison, M.; Herdtweck, E.; Priermeier, T.; Kiprof, P. *J. Organomet. Chem.* **1995**, *498*, 1.
(15) Senkan, S. *Angew. Chem. Int. Ed.* **2001**, *40*, 312.
(16) Cundari, T. R.; Deng, J.; Pop, H. F.; Sârbu, C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1052.
(17) Lipkowitz, K. B.; Pradhan, M. *J. Org. Chem.* **2003**, *68*, 4648.
(18) Cruz, V.; Ramos, J.; Muñoz-Escalona, A.; Lafuente, P.; Peña, B.; Martinez-Salazar, J. *Polymer* **2004**, *45*, 2061.
(19) Cruz, V. L.; Ramos, J.; Martinez, S.; Muñoz-Escalona, A.; Martinez-Salazar, J. *Organometallics* **2005**, *24*, 5095.
(20) Lü, Q.; Yu, R.; Shen, G. *J. Mol. Catal. A* **2003**, *198*, 9.
(21) Occhipinti, G.; Bjørsvik, H.-R.; Jensen, V. R. *J. Am. Chem. Soc.* **2006**, *128*, 6952.
(22) Orpen, A. G. *Acta Crystallogr. B* **2002**, *58*, 398.
(23) Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A. *Chem. Eur. J.* **2006**, *12*, 291.

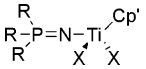


Figure 1. General template for the catalytic system studied. Cp′ indicates any species with a cyclopentadienyl ring; X may be Cl, Me, or CH_2SiMe_3.

the use of these categories in proposing new catalysts was not discussed. The first, more specific definition of similarity is very familiar to synthetic chemists, and therefore, in a sense, this database mining procedure is merely an automated means to propose ligands that could be arrived at by any chemist with a pencil and a pad of paper. The second measure of similarity, however, has been used with the goal of proposing novel ligands unfamiliar in shape, but with pleasantly familiar properties, such as high activity.

In order to use these three drug discovery tools—2D- and 3D-QSAR and database mining—to rationally design organometallic catalysts, it will be beneficial, and indeed necessary in 3D-QSAR, to study series of related compounds where a central motif is preserved, with changes in activity brought about by modifications around, rather than within, this core. This philosophy is no different from traditional approaches used to improve catalysts,[24−28] where many compounds are largely identical, save for the addition of, e.g., steric bulk or an electron-withdrawing group, to investigate the effects of sterics or electrostatics, respectively. Indeed, the computational drug discovery tools explored in the current and similar studies[17−21] merely provide a systematic manner in which to rationalize the effects of structural variation, with the added benefit of providing insight into complicated, multivariate effects by using statistical techniques such as principal component analysis (PCA) or partial least-squares regression (PLSR).[29]

The central structural motif chosen in the current work to test the utility of computational drug discovery techniques in catalyst design is the cyclopentadienyl−titanium phosphinimide (Ti−N═P) complexes of Stephan et al.,[30] illustrated schematically in Figure 1 and described in detail in Table 1. The catalysts in this system are not only effective ethylene polymerization catalysts but also possess three potential modification sites, affording the flexibility necessary to successfully engineer improved catalysts. From the experimental work,[30] only the titanium-based species whose activities were measured under the "MAO/a" co-catalyst conditions were included to eliminate uncertainty due to variation in experimental conditions. In principle, any property can be related to structure with a 2D- or 3D-QSPR approach, such as the polydispersity index, the average weight of polymer produced, or the amount of enantiomeric excess produced.[17,31] However, in the current work, the property of interest to be correlated with structural variation is activity, defined as the amount of polymer, in g/mmol/h/atm, produced by each catalyst.

Table 1. Details of the Catalysts Modeled[a]

compound	X	Cp′	R	activity[e]
10	Cl	C_5H_5	Et	13
11			Cyc	42
12			^{i}Pr	49
13			^{i}Bu	652
14			Ph	34
15			*p*-MeC_6H_4	35
16			*p*-$CF_3C_6H_4$	31
17			*p*-FC_6H_4	34
18			*p*-$MeOC_6H_4$	47
19		$C_5H_4SiMe_3$	^{i}Pr	16
20			^{i}Bu	494
21		C_5Me_5	^{i}Pr	30
22			^{i}Bu	1400
23		Indenyl	^{i}Pr	28
24			^{i}Bu	425
25		$C_5H_4(^t$Bu)	Cyc	46
26			^{i}Pr	16
27			^{t}Bu	881
28		$C_5H_4(^n$Bu)	^{t}Bu	2000
44	Me	$C_5H_4(^t$Bu)	^{t}Bu	853
62	CH_2SiMe_3	C_5H_5	Ph	765
$CpTiCl_3$[b]	Cl		Cl	9[f]
CGC[c]		$C_5Me_4SiMe_2$	^{t}Bu[d]	630

[a] Details and numbering from ref 30. [b] This structure is as written and does not contain the N═P motif. [c] The constrained geometry catalyst, $C_5Me_4SiMe_2N(^t$Bu)$TiCl_2$; see ref 32. [d] This is a single ^{t}Bu group, rather than $P(^t$Bu$)_3$. [e] In g/mmol/h/atm. [f] Reported as <10.

In this work, the catalytic system described in Table 1 was used as a testing ground to investigate the possibility of applying computational drug discovery tools to select ligands, from among a large database, likely to yield highly active organometallic catalysts. The software implementing these drug discovery approaches was used in a "black box" fashion—that is, without any code modification. In addition, all chosen software is freely available, at least on a trial basis, and fairly straightforward to use. These characteristics are all desirable in a field as new (see above) as the application of drug discovery approaches, originally intended for exclusively organic systems, to organometallic systems. Lessons learned from this black box approach can be used to guide future efforts to refine the specific techniques used to mix the two disciplines of organometallic catalysis and organic drug design.

This work is organized as follows. After a description of the technical details of the pieces of software chosen, information is given concerning the construction of models, using 3D-QSAR, to rationalize the catalyst activities given in Table 1; a similar treatment for models constructed with 2D-QSAR follows. Next, the specifics of the database mining

(24) Hou, Z.; Zhang, Y.; Tezuka, H.; Xie, P.; Tardif, O.; Koizumi, T.; Yamazaki, H.; Wakatsuki, Y. *J. Am. Chem. Soc.* **2000**, *122*, 10533.
(25) Doctrow, S. R.; Huffman, K.; Marcus, C. B.; Tocco, G.; Malfroy, E.; Adinolfi, C. A.; Kruk, H.; Baker, K.; Lazarowych, N.; Mascarenhas, J.; Malfroy, B. *J. Med. Chem.* **2002**, *45*, 4549.
(26) Groysman, S.; Tshuva, E. Y.; Goldberg, I.; Kol, M.; Goldschmidt, Z.; Shuster, M. *Organometallics* **2004**, *23*, 5291.
(27) Speiser, F.; Braunstein, P.; Saussine, L. *Acc. Chem. Res.* **2005**, *38*, 784.
(28) Wölfle, H.; Kopacka, H.; Wurst, K.; Ongania, K.-H.; Görtz, H.-H.; Preishuber-Pflügl, P.; Bildstein, B. *J. Organomet. Chem.* **2006**, *691*, 1197.
(29) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1.
(30) Stephan, D. W.; Stewart, J. C.; Guérin, F.; Courtenay, S.; Kickham, J.; Hollink, E.; Beddie, C.; Hoskin, A.; Graham, T.; Wei, P.; Spence, R. E. v. H.; Xu, W.; Koch, L.; Gao, X.; Harrison, D. G. *Organometallics* **2003**, *22*, 1937.
(31) Hoogenraad, M.; Klaus, G. M.; Elders, N.; Hooijschuur, S. M.; McKay, B.; Smith, A. A.; Damen, E. W. P. *Tetrahedron: Asymmetry* **2004**, *15*, 519.
(32) Woo, T. K.; Margl, P. M.; Lohrenz, J. C. W.; Blöchl, P. E.; Ziegler, T. *J. Am. Chem. Soc.* **1996**, *118*, 13021.

procedures are given, as are resulting ligands proposed to yield highly active catalysts when grafted onto the Ti−N═P framework. The activities of new catalysts created in this fashion are predicted using the 3D- and 2D-QSAR models in the next section. On the basis of these predicted activities, a discussion is offered about the strengths and drawbacks of the three drug discovery tools in the context of organometallic catalyst design, and conclusions, particularly about the future of this line of research in general, are offered.

II. Computational Methods

Geometry Optimization. Given the lack of X-ray crystal structures for many of the catalysts given in Table 1, the geometries of all structures were optimized using DFT as implemented in NWChem 4.7,[33] specifically the PW91 functional[34] in conjunction with a LANL2DZ ECP basis set for Ti, Cl, Si, and P atoms and with a 6-311 g+* basis set, as implemented in NWChem, for all other atoms. The fine keyword was invoked to define grid size; all other options were left at their default values. The effect of the choice of functional and basis set on constructed structure−activity model accuracy is beyond the scope of this work, although results (not shown) indicate that calculated predicted activities are largely insensitive to variations in geometry on the order of 0.2 Å. PW91 has been shown[35] to generally provide excellent geometries for organometallic systems. Moreover, the intention of geometry optimization prior to model construction is consistency, rather than strict agreement with experimentally characterized geometries.

3D-QSAR. The free program SOMFA (Self-Organizing Molecular Field Analysis)[36] was used for construction of 3D-QSAR models. The catalysts in Table 1, as well as all newly proposed catalysts, were centered in a cube measuring 50 Å per side. Two fields can be evaluated in SOMFA: sterics and electrostatics. Attempts to rationalize electrostatic patterns, based on calculated Mulliken atomic charges, did not prove successful, and therefore, only the steric field was included in this study. The original experimental researchers[30] of the Ti−N═P system studied did note that patterns in molecular shape, rather than electrostatics, seem to be largely responsible for determining activity. For other systems where partial charges are likely to be influential, predicted activities could benefit from using this SOMFA field. Returning to the present system, the steric field was evaluated using a grid with 0.25 Å resolution. Proper alignment of molecules is a well-known prerequisite for successful 3D-QSAR models. Therefore, all molecules were aligned, using **10** as a template, with the shareware program VEGA.[37]

2D-QSAR. The molecular descriptors that serve as the input for the construction of 2D-QSAR models were calculated using the Web-based program E-Dragon,[38] which allows for the evaluation of over 1600 properties. File format conversions between the Cartesian coordinates obtained with NWChem and the Structure Data Format (SDF)[39] required by E-Dragon were performed with the shareware program Mol2Mol.[40] Following descriptor calculation, models were generated using the PLSR capabilities of the demo version of The Unscrambler 9.6.[41] In order to build an effective 2D-QSAR model, it is advisable to first pare down the list of descriptors to remove redundancies, collinearity, and properties that do not vary greatly across the compound set.[42] To this end, models were built from three different sets of calculated descriptors: the complete set, the set after treatment with a Web-based unsupervised forward selection (UFS) algorithm,[43] and the set after treatment with a Web-based genetic algorithm (GA).[44] Prior to PLSR, all descriptors were mean-centered and normalized to prevent descriptors with numerically large values from dominating the 2D-QSAR model.

Database Mining. In order to propose new ligands for replacing the R and Cp′ groups of the Ti−N═P framework, two database mining approaches, each based on a different measure of similarity, were used. For the first, a structural similarity search, the implementation of the freely available Chemmine program was used.[45] The "atom pairs" search method gave all of the results of the "atom sequence" search method and more and was therefore chosen as the default searching method. Structural similarity in this program is measured by the Tanimoto coefficient,[46] which ranges from 0 (no similarity) to 1 (identical). The minimum cutoff for this coefficient was determined ad hoc to give a sufficient number of hits. Hit lists were further culled on the basis of structural considerations (e.g., the presence of an aromatic pentacycle for proposed Cp′-type ligands). The second database mining method, based on a more general measure of similarity, will be described in greater detail later but, to summarize, involves the calculation (again using E-Dragon) of descriptors for all structures in a database, the identification (using PLSR in Unscrambler) of the 2D-QSAR descriptors most strongly correlated with activity, and the selection of ligands in the database with high values for these descriptors.

Compound Databases. Two databases were mined for ligands that may potentially lead to improved catalysts when attached to the Ti−N═P framework. These databases are freely available from the Development Therapeutics Program of the National Cancer Institute website[47] (indeed, quantities of the actual compounds themselves are also available). The main database in this work, termed Plated, consists of over 140 000 compounds that have been evaluated as potential anti-HIV and anti-cancer agents. A subset, termed Diversity, contains 1990 compounds and serves as a representative sample of the structural diversity present in the larger

(33) (a) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671. (b) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1993**, *48*, 4978.

(34) *High Performance Computational Chemistry Group, "NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.7"*; Pacific Northwest National Laboratory: Richland, WA, 2005.

(35) Drummond, M. L. Ph.D. Disseration, The Ohio State University, 2005.

(36) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. *J. Med. Chem.* **1999**, *42*, 573.

(37) Pedretti, A.; Villa, L.; Vistoli, G. *J. Comput. Aided Mol. Des.* **2004**, *18*, 167.

(38) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. *J. Comput. Aided Mol. Des.* **2005**, *19*, 453.

(39) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244.

(40) Gunda, T. E. *Mol2Mol*, version 5.5; University of Debrecen: Debrecen, Hungary, 2006.

(41) *The Unscrambler*, version 9.6; Camo: Trondheim, Norway, 2006.

(42) Kubinyi, H. In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., Eds.; Dekker Inc.: New York, 2003; p 539.

(43) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160.

(44) Palyulin, V. A.; Radchenko, E. V.; Baranova, O. D.; Oliferenko, A. A.; Zefirov, N. S. In *EuroQSAR2002 Designing Drugs and Crop Protectants: Processes, Problems, and Solutions*; Ford, M., Livingstone, D., Dearden, J., van de Waterbeemd, H., Eds.; Blackwell Publishing: Oxford, 2003; p 188.

(45) Girke, T.; Cheng, L.-C.; Raikehl, N. *Plant Physiol.* **2005**, *138*, 573.

(46) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.

(47) http://dtp.nci.nih.gov/index.html.

Plated set. The choice of these databases was based solely on their availability; any database can of course be mined.

Statistical Metrics. Standard statistical means, detailed in the Supporting Information, were used to evaluate all QSAR models constructed.

III. Results

3D-QSAR. As 3D-QSAR requires compound geometries as input and as not all catalysts in Table 1 were characterized by X-ray crystallography, the geometry of each compound was calculated with DFT. The accuracy of this procedure compared to experiment can be evaluated for those catalysts that do have crystal structures, namely **11**−**13**, **21**−**23**, **27**, and **28**. For these compounds, bond distances, specifically the Ti−Cl, Ti−N, and N−P distances, are calculated with an RMSD of 0.052 Å and a maximum error of 0.090 Å. Similarly, bond angles (Cl−Ti−Cl, N−Ti−Cl, Ti−N−P, and N−Ti−$Cp_{centroid}$) are calculated with an RMSD of 3.2° and a maximum error of 10.5°. Whether these calculations are of "sufficient" precision is somewhat subjective, but results (not shown) indicate that calculated predicted activities are largely insensitive to variations in geometry on the order of 0.2 Å.

Nevertheless, despite any fortuitously small influence of inaccurately calculated geometries, it is generally advisable that all structures upon which a 3D-QSAR model is based be aligned as closely as possible. Testing, detailed in the Supporting Information, revealed that alignment of $Cp_{centroid}$, Ti, and N points between the various catalysts gave the closest overlap. In addition, this alignment yielded the best statistical measurements of predicted activities ($r^2 = 0.666$, q^2-CV = 0.216), and therefore, this alignment scheme was chosen as the basis for all 3D-QSAR models discussed. A visual representation of the 3D space covered by this model is shown in Figure 2.

Following alignment, the catalysts of Table 1 were divided into training sets, used to build the 3D-QSAR model, and test sets, used to evaluate the predictive capabilities of the model for species not included in the model itself. The test sets are described in Table 2, as are the statistical metrics for each. Sets 1−5 were each constructed by excluding five randomly selected catalysts prior to model construction, whereas Set 6 consists of the exclusion of the two structures that do not fit the Ti−N═P framework, **CGC** and **$CpTiCl_3$**, as well as the two structures where the X groups are not Cl. A few observations about the models constructed from the training/test set splits described in Table 2 can be made. In general, these models are fit as well as the model constructed above for all species ($r^2 = 0.666$), and, indeed, models based on Sets 3 and 4 seem to be even better. More noticeably, q^2 values are greatly improved compared to above (q^2-CV = 0.216), particularly for the models based on Sets 3 and 6.

The significance of proper training/test splitting has previously been discussed by Kubinyi.[48] An exhaustive evaluation of models constructed following exclusion of all possible four- and five-membered test sets (8855 and 33 649 distinct splits, respectively), as suggested,[48] cannot be performed in a timely manner with SOMFA. One alternative is to use a categorization approach such a PCA[23] in order to identify and properly address catalysts with truly outlying activities. Instead, the approach chosen in this work is to use all six models described in Table 2, with the modifications discussed below, to investigate how much predicted activities for new catalysts vary on the basis of the training/test set split choice.

Initial use and testing of the models in Table 2 revealed that one effect not yet addressed often has a profound effect on the accuracy of predicted activity, namely the fluxionality of the Cp′ ligand, often termed "ring whizzing". As illustrated in Figure 2, only one stereoisomer of each species in Table 1 is incorporated into the models of Table 2. This is most obvious for **28**, where the ^{n}Bu group can be seen in the lower right of Figure 2. According to Table 1, **28** is the most active catalyst, obviously due to the presence of this alkyl substituent on the Cp ring. However, if a "new" catalyst were formed by rotating the Cp′ ligand of **28** in the plane of Figure 2, as is possible to at least some degree, the activity of this "new" catalyst predicted by any of the 3D-QSAR models in Table 2 would fall far short of the experimental activity of 2000, because these models reflect the fact that an ^{n}Bu group

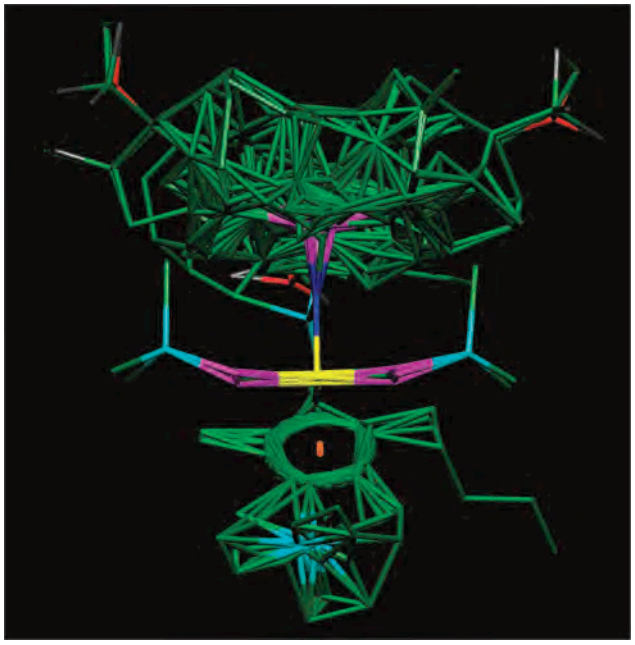


Figure 2. Alignment of the catalysts of Table 1, aligned at $Cp_{centroid}$, Ti, and N. The coloring scheme is as follows. $Cp_{centroid}$, orange; C, green; Ti, yellow; Cl, purple (to the left and right of Ti); Si, light blue; N, dark blue; P, purple (above Ti); O, red; F, gray.

Table 2. Details on the Test Sets Used for the 3D-QSAR Models

set	compounds excluded	r^2	q^2-Test	RMSD
1	**12**, **14**, **25**, **62**, CGC	0.669	0.373	265
2	**13**, **14**, **22**, **25**, **62**	0.657	0.507	403
3	**13**, **15**, **18**, **22**, **24**	0.777	0.626	326
4	**14**, **17**, **20**, **27**, **44**	0.765	0.392	304
5	**19**, **20**, **25**, **44**, $CpTiCl_3$	0.638	0.485	256
6	**44**, **62**, $CpTiCl_3$, CGC	0.663	0.634	245

(48) Kubinyi, H. In *QSAR & Molecular Modelling in Rational Design of Bioactive Molecules*; Proceedings of the 15th European Symposium on QSAR & Molecular Modelling; Istanbul, Turkey, 2004; Sener, E. A., Yalcin, I., Eds.; CADDD Society: Ankara, Turkey, 2006; p 30.

Table 3. Performance of the Fluxional 3D-QSAR Models

set[a]	r^2	q^2-Test	RMSD
1	0.826	0.527	326
2	0.816	0.547	349
3	0.862	0.403	333
4	0.842	0.777	207
5	0.827	0.790	178
6	0.817	0.556	304

[a] See Table 2 for set definition.

only in the spatial position shown in Figure 2 leads to a high activity. In order to rectify this shortcoming, the asymmetric Cp′ ligands of the catalyst in Table 1 were rotated about the Ti−X axis in 72° increments. The five stereoisomers thus produced for each species were then visually inspected for possible steric clashes with the rest of the molecule. All new stereoisomers without steric clashes were assigned the activities listed in Table 1, and new 3D-QSAR models were built from all acceptable rotamers for each catalyst, using the same training/test set splits described in Table 2. The statistical measures of these new fluxional models are given in Table 3. While all models show improved internal fits (r^2), surprisingly the predictivity of the test set is worsened for Sets 3 and 6, as measured by comparing the respective q^2-Test values with the q^2-Test values of Table 2. However, all other models show significant improvement with this approach, indicating its utility. Further improved models could likely be produced if, as has been done elsewhere,[49] data from molecular dynamics were incorporated to provide a measure of fluxionality more accurate than the evenly weighted contribution of five rotamers. Nevertheless, this approximation to a wholly dynamic 3D-QSAR approach is the most accurate procedure considered in this work, and therefore, the models described in Table 3 are deemed the final 3D-QSAR models and will be used to evaluate the activities of newly proposed catalysts.

2D-QSAR. In evaluating the applicability of computational drug discovery tools to organometallic catalyst discovery, one of the challenges is that the former were designed for druglike molecules, almost none of which contain metal atoms. This difference was highlighted immediately in the development of 2D-QSAR models for new organometallic catalysts, as the freely available program chosen to calculate molecular descriptors, E-Dragon,[38] uses a file format (SDF) that does not contain a definition for titanium. Given that this program calculates over 1600 descriptors using a variety of formulas, modifying the code to accommodate Ti would likely present a significant challenge. In addition, as this is an early study of the use of drug discovery tools for organometallic catalyst design, it was decided to apply all software in an unmodified, black box fashion to serve as a foundation upon which future efforts at mixing the two disciplines could be based. Thus, it was decided to empirically determine which element described by E-Dragon is closest to Ti, as measured by the accuracy of predicted activities. Thus, 2D-QSAR models were generated using the training/test splits of Table 2 with Ti replaced by 17 different elements. The best results were found for C and Ag, with the best RMSDs found for the test sets no. 4−111 and 148, respectively. The transition metals Cr, Au, Co, Fe, and Ni were less acceptable, with RMSDs ranging from 182 to 200; other elements yielded even poorer results. Therefore, for all 2D-QSAR models described below, Ti was replaced with C. The effect of this somewhat drastic change will be judged on the basis of the accuracy of predicted activities, both for test sets and for newly proposed catalysts.

Table 4. Final 2D-QSAR Models Produced

set	no. of LVs	r^2	q^2-Test	RMSD
1	1	0.483	0.271	313
2	7	1.000	0.066	522
3	10	0.999	0.485	383
4	6	0.965	0.919	111
5	1	0.643	−3.994	796
6	9	0.928	−2.933	803

Another challenge in 2D-QSAR model generation is choosing which descriptors should be correlated with activity. To prevent overfitting, where too many variables are used to build a model, it is necessary to choose descriptors that are neither redundant nor collinear. In order to eliminate useless or flawed descriptors, two algorithms were applied (see above)—an unsupervised forward selection (UFS) approach and a genetic algorithm (GA). As a control, models were also generated without any descriptor elimination. Activities for the catalysts in the test sets predicted by the 2D-QSAR models following application of the UFS showed RMSDs increased by a factor of 1−20 compared to models built with the entire descriptor set. In other words, the UFS procedure was too ambitious in paring down the descriptor list, leading to very poor predicted activities. In contrast, the models generated following application of the GA showed, at worst, activities comparable to those predicted with the entire descriptor set. For all but Set 6, the rmsd of predicted activities of GA-based models were improved, with the best improvement, 41%, found for Set 4. This finding clearly illustrates the benefits of avoiding an overfitted model, and thus, for all 2D-QSAR models developed below, the descriptor set pared by the GA was used as input for model generation with PLSR.

In PLSR, potentially correlated variables (i.e., the set of descriptors calculated by E-Dragon and pared down by the GA) are condensed into a small number of orthogonal variables, often called latent variables (LVs) or PLS factors, consisting of weighted combinations of the E-Dragon descriptors. The number of LVs chosen should explain as much variance in the data with as few LVs as possible. In other words, it is seldom advisable to add another LV if only a few additional percent of the variance in the data is explained, as this improvement most often represents modifications to the model to describe only a single compound, and therefore, the model will likely be inapplicable to compounds outside of the training set. With this caveat, Table 4 lists the final 2D-QSAR models constructed, using the test set definitions of Table 2, with Ti replaced by C, and with the descriptor set culled by the GA.

(49) Melville, J. L.; Lovelock, K. R. J.; Wilson, C.; Allbutt, B.; Burke, E. K.; Lygo, B.; Hirst, J. D. *J. Chem. Inf. Model.* **2005**, *45*, 971.

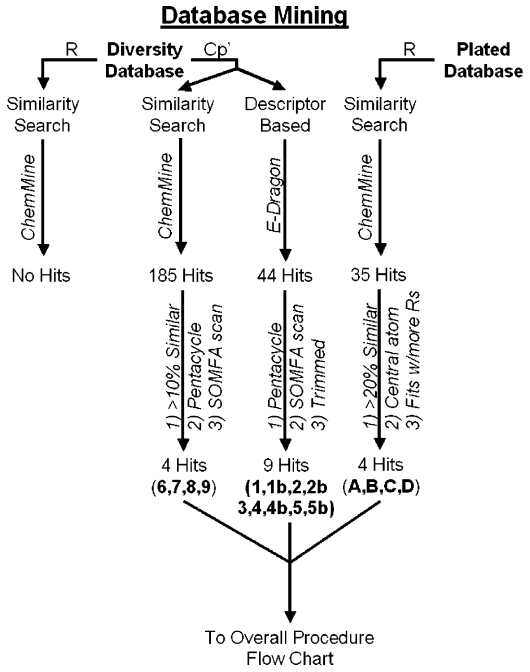


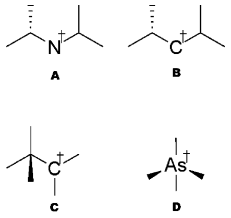
Figure 3. Summary of the database mining procedures used. See text for details.

Four of the models shown in Table 4—2, 3, 4, and 6—are fit extremely well, with r^2 values of over 0.9. However, examination of the q^2-Test metrics indicates that, despite the application of the GA and the generation of orthogonal LVs, most of these fits can be regarded as overfitted. Indeed, the negative values of q^2-Test for Sets 5 and 6 indicate that the predicted activities for the test set compounds are worse than a "model" where each unknown, predicted activity is simply set to the average activity of the compounds of the test set. Only the model built from training/test set split 4 indicates both a good fit of the training set ($r^2 = 0.965$) and good predictions of the test set (q^2-Test = 0.919). Scrambling the data prior to model construction, as described in the Supporting Information, further reveals the statistical validity of Model 4.

Database Mining. As stated in the Introduction, the critical hurdle in designing new catalysts with potentially beneficial properties is identifying the few useful ligands among the far larger set of useless ligands. To attempt to overcome this obstacle, two different mining methods were employed, based on two different definitions of "similarity." In addition, two databases were mined and two ligand types were sought after—Cp′ and R (Figure 1). The overall procedure is summarized in Figure 3 and will be described in detail below.

Similarity Search. The first database mining approach utilized a similarity search algorithm, which simply finds ligands in the compound database possessing a connectivity similar to the search target. Perusal of Table 1 suggests that ^{t}Bu is the optimum R ligand (cf. **19** and **20**, **21** and **22**, etc.), and therefore, the two compound databases were mined for structures similar to ^{t}Bu. In the Diversity database, no hits of >20% similarity resulted, and thus, the far larger Plated database was mined. This search resulted in 35 hits with >20% similarity. However, manual inspection of these ligands revealed that the majority of these hits, while containing a moiety similar to a ^{t}Bu group, were quite large, thereby preventing the attachment of three such ligands to the single phosphorus of Ti−N═P. (The possibility of replacing three individual R groups with a single tridentate R group was not considered.) Of the 35 hits, four ligands of a size comparable to ^{t}Bu were identified, each possessing a central atom with a single hydrogen atom that could be removed prior to ligation to the P of Ti−N═P. These four ligands are shown in Chart 1. At first glance, these ligands are quite simple and could be readily proposed as structurally similar replacements for ^{t}Bu. However, the other ligand searches will serve to demonstrate the necessity for a computational screen to obtain diverse and novel ligands, leading to potentially high activity catalysts.

Chart 1. Newly Proposed R-type Ligands[a]

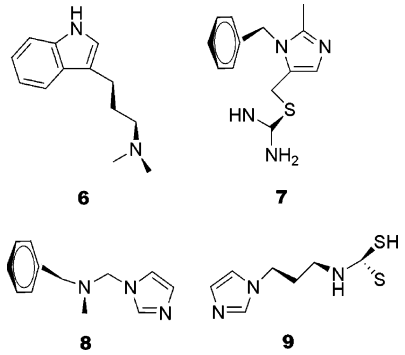


[a] The cross indicates the point of attachment to P(═N−Ti).

The similarity search algorithm was also used to find additional Cp′-type ligands. The Cp′ portions of both **22**, C_5Me_5, and **28**, $C_5H_4(^n\text{Bu})$, were the search targets, as catalysts with these ligands possess activities substantially above the norm (Table 1). When the Diversity database was searched for structures similar to C_5Me_5, 12 hits resulted with similarity >10%, three of which contained an aromatic pentacycle similar to Cp. When the Diversity database was mined for analogues of $C_5H_4(^n\text{Bu})$, 164 hits were produced with >10% similarity, 18 of which possessed an aromatic pentacycle moiety, leading to a total of 21 aromatic pentacycles in the Diversity database similar to either C_5Me_5 or $C_5H_4(^n\text{Bu})$. This list was pared down further with fluxional Model 5 of Table 3, which was used to predict the activities of newly formed catalysts consisting of ^{t}Bu groups for R and the 21 aromatic pentacycles in place of Cp′. Four Cp′-type ligands led to catalysts with high predicted activities, and these are the final hits, shown in Chart 2, of the similarity-based Cp′ search of the Diversity database. As the similarity search Diversity database yielded 185 hits, or roughly 10% of the entire database, the far larger Plated database was not mined in this fashion.

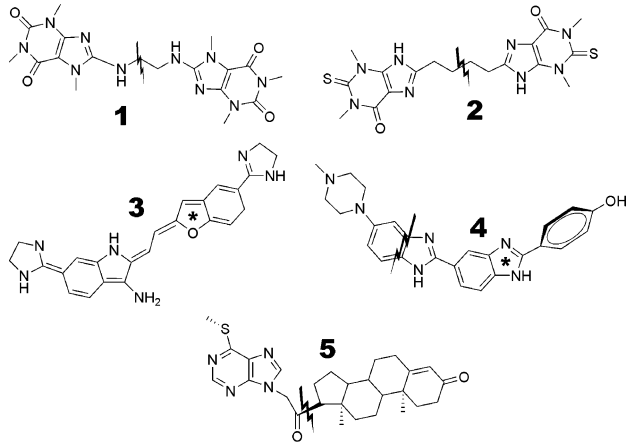
Descriptor-Based Search. The first step in performing what is termed the descriptor-based search is the calculation of descriptors for an entire compound database. The freely available program chosen for descriptor evaluation in this work, E-Dragon, can only calculate the descriptors of 150 molecules at a time. Thus, databases containing more than 150 compounds require multiple runs that, to the best of our knowledge, cannot be automated. While this restriction is merely impractical for the smaller Diversity database, requiring 14 separate submissions, it is very restrictive for

Chart 2. Newly Proposed Cp′-type Ligands, from a Similarity Search of the Diversity Database[a]



[a] Ligation to the Ti(−N═P) is through the aromatic pentacycle.

Chart 3. Newly Proposed Cp′-type Ligands, from a Descriptor-Based Search of the Diversity Database[a]



[a] The jagged line indicates removal of the smaller part (or the aliphatic part, in the case of **5**), leading to the creation of **1b**, **2b**, **4b**, and **5b**. Ligation to Ti is through an aromatic pentacycle, indicated by an asterisk when this is ambiguous.

the 140 000-member Plated database. Therefore, only the Diversity database was mined in the fashion described below.

The key to this procedure is the identification of the descriptors for each ligand type that most strongly correlate with high predicted activities. While it is possible to simply investigate the PLSR results of the models shown in Table 4, the goal is to search for new R-type and Cp′-type ligands independently from one another; the models in Table 4 predict activities based on the simultaneous contribution of both ligand types. Therefore, new heuristic 2D-QSAR models were constructed by taking only the Cp′ ligand geometry of the catalysts in Table 1 where R = ^{t}Bu, that is, a lone Cp group from **13**, a $C_5H_4SiMe_3$ group from **20**, and similarly for **22**, **24**, **27**, **28**, and **44**. As each catalyst contains the same R ligand, the variation in activity can be attributed solely to the choice of Cp′ group. Therefore, the activity of each individual Cp′-type ligand was set to the value listed in Table 1 for the entire catalyst. From this Cp′-only input, 2D-QSAR models were built as before and the PLSR data were inspected to identify the 10 descriptors most important in determining activity. For each of these descriptors, standard deviations were calculated across all compounds in the Diversity database. Next, the compound with the highest overall value was tabulated for each of the 10 descriptors, as were all compounds possessing descriptors within 2 SDs of each maximum value. Considerable redundancy among the hit lists for the 10 descriptors resulted in a total of 44 compounds in this table, 11 of which contained an aromatic pentacycle. This list was pared down somewhat by using the 3D-QSAR Model 5 of Table 3, with R set to ^{t}Bu and Cp′ set to each of the 11 hits. Five ligands outperformed the other seven, and these are given in Chart 3. These ligands are all substantially larger than even the largest Cp′-type ligand listed in Table 1. Thus, to investigate the effect of their larger sizes, new ligands were also proposed consisting of the ligands in Chart 3 truncated at the point indicated by the jagged line. Ligand **3** was not trimmed in this fashion, as it was initially considered in two forms, ligated to Ti via the furan ring (as shown) or via the central pyrrole ring; the latter was revealed by the 3D-QSAR screen mentioned above to lead to poorer activity and was thus removed from the hit list. Overall, the nine ligands described by Chart 3 represent the final output of the descriptor-based screen of the Diversity database for Cp′-type ligands. An attempt to locate new R-type ligands using this approach produced no ligands of a size to attach to the P atom of Ti−N═P along with two other R-type ligands.

The database mining procedures described above could certainly be tailored, depending on the desired result. In this study, improvements at both the R and Cp′ sites were sought. Thus, because newly proposed ligands at one site must be crossed with all proposed ligands at the other site, hit lists were aggressively trimmed multiple times to avoid a combinatorial explosion. This approach is especially necessitated by our preferred 3D-QSAR approach, where up to five rotamers must be created for each Cp′-type ligand to address fluxionality. However, if only a single ligand site were of interest, or if an entirely static 3D-QSAR model were employed to predict catalyst activities, fewer ligands could be removed from database mining results.

To summarize, five R-type ligands—**A**−**D** of Chart 1, as well as ^{t}Bu to serve as a control—and 14 Cp′-type ligands—**1**, **1b**, **2**, **2b**, **3**, **4**, **4b**, **5**, and **5b** in Chart 3, as well as **6**−**9** in Chart 2, and the control of Cp—resulted from the database mining procedures of Figure 3, leading to 69 new catalysts when crossed (^{t}Bu/Cp cross is simply **13**). The predicted activities of these new catalysts, using both the 3D- and 2D-QSAR approaches described above, can now be discussed. The entire procedure used to propose new catalysts and predict their activities is summarized in Figure 4.

Predicted Activities of New Catalysts. 3D-QSAR. Figure 5 shows the predicted activities, using the six 3D-QSAR models given in Table 3, for the Ti−N═P-type catalysts formed from the 13 newly proposed Cp′-type ligands with R = ^{t}Bu. The predicted activities in this figure are fairly consistent regardless of the specific fluxional 3D-QSAR model used; similar consensus is noted in similar plots where R = **A**−**D**. The standard deviations across the six models in Figure 5 range from 30 (for ligands **6** and **9**) to 66 (for ligand **5**). The model indicated in Table 3 as possessing the best external predictivity, Model 5, tends to yield the highest

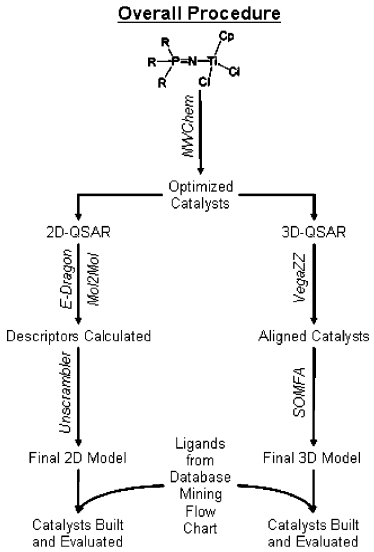


Figure 4. Summary of the overall procedure used to propose new catalysts and predict their activities.

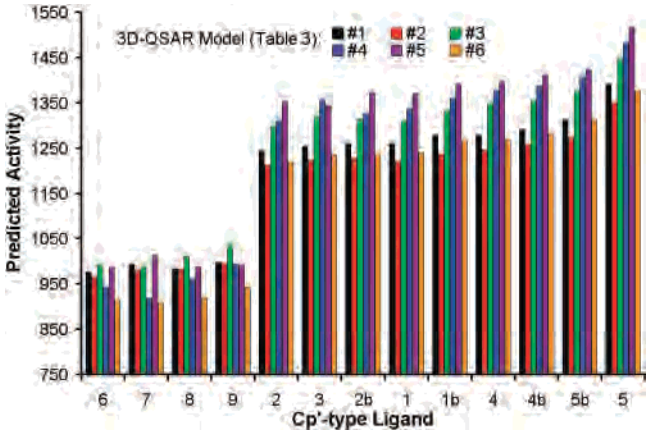


Figure 5. 3D-QSAR predicted activities for the newly proposed catalysts where R = ^{t}Bu.

Table 5. Predicted Activities for All New Catalysts, Using 3D-QSAR Model 5 of Table 3

	R				
Cp′	**A**(^{t}Bu)$_2$	**B**(^{t}Bu)$_2$	**C**$_3$	**D**$_2$(^{t}Bu)	(^{t}Bu)$_3$
1	776	794	852	898	1370
1b	795	813	871	915	1391
2	774	794	855	901	1353
2b	794	814	875	921	1373
3	778	790	824	898	1345
4	791	809	866	927	1397
4b	797	814	872	934	1411
5	916	934	994	1036	1518
5b	822	842	904	949	1425
6	893	911	971	1018	988
7	926	943	1004	1047	1013
8	901	919	982	1019	989
9	906	924	993	1030	991
Cp	812	829	892	933	901[a]

[a] This is the predicted activity of **13** (actual = 652).

predicted activities. The activities predicted with this model for all newly proposed catalysts are given in Table 5.

Starting with the last row of this table, it is seen that only ligand **D** represents an improvement in the R-type ligand over ^{t}Bu, and this improvement is marginal. Moreover, while no newly proposed catalyst outperforms the best original catalyst, **28**, which has a predicted activity of 1634 with Model 5 of Table 3 (actual = 2000), the Cp′-type ligands arrived at through the descriptor-based mining process (**1**−**5**) all lead to catalysts that outperform the *second best* original catalyst, **22** (predicted = 1098, actual = 1400), when R = ^{t}Bu. Finally, truncating the especially large Cp′-type ligands does not have a large effect on predicted catalytic activities.

Table 6. Predicted Activities for All New Catalysts, Using 2D-QSAR Model 4 of Table 4

	R				
Cp′	**A**(^{t}Bu)$_2$	**B**(^{t}Bu)$_2$	**C**$_3$	**D**$_2$(^{t}Bu)	(^{t}Bu)$_3$
1	73 624	83 288	99 368	94 088	73 254
1b	2425	2456	3747	3386	2556
2	66 770	59 010	71 444	61 784	54 798
2b	2785	2874	6695	3973	2961
3	82 615	95 615	119 986	90 465	85 448
4	65 040	61 862	93 850	98 158	70 285
4b	7934	8069	27299	6921	7286
5	113 275	107 750	128 900	129 770	100 825
5b	2350	2400	3110	3253	2486
6	8436	8526	18554	8489	6932
7	10 128	10 310	24 134	12 666	8900
8	5639	5669	21495	7454	5423
9	22 921	22 666	36 501	42 465	21 362
Cp	2237	2276	2832	2918	792[a]

[a] This is the predicted activity of **13** (actual = 652).

Table 7. The 10 Most Active Catalysts as Predicted by Model 4 of Table 4 and the Ranking of These Catalysts Predicted by the Other 2D-QSAR Models

		rankings within each 2D-QSAR model					
R	Cp′	4	1	2	3	5	6
D	**5**	1	3	2	3	2	1
C	**5**	2	1	1	1	1	2
C	**3**	3	10	3	5	3	3
A	**5**	4	6	4	2	4	4
B	**5**	5	7	5	4	5	5
(^{t}Bu)$_3$	**5**	6	12	6	8	7	8
C	**1**	7	4	7	7	6	7
D	**4**	8	15	8	12	9	6
B	**3**	9	23	11	15	11	9
D	**1**	10	2	10	9	8	10

2D-QSAR. In stark contrast to the 3D-QSAR results, the different 2D-QSAR models are most certainly not in quantitative agreement with each other. Moreover, the predicted activities (Table 6) of new catalysts using the best 2D-QSAR model, no. 4 of Table 4, cover a range far broader than that indicated by 3D-QSAR. Indeed, the most successful catalyst as measured by 2D-QSAR has a predicted activity over 65 times greater than the best experimentally measured catalyst activity in Table 1, which, absent an experimental confirmation, seems to give reason for skepticism about the quantitative accuracy of the results in Table 6. Additionally, a comparison of the full Cp′-type ligands **1**, **2**, **4**, and **5** with their truncated partners reveal that, unlike the 3D-QSAR results, the trimmed ligands have severely lower predicted activities compared to their untrimmed counterparts. The most extreme example is the contrast between **A**(^{t}Bu)$_2$-**5** and **A**(^{t}Bu)$_2$-**5b**, where the catalyst built from the truncated ligand shows a predicted activity of only 2.1% of the catalyst built from the full ligand. It seems unlikely that the mere elimination of an admittedly sizable side group could have

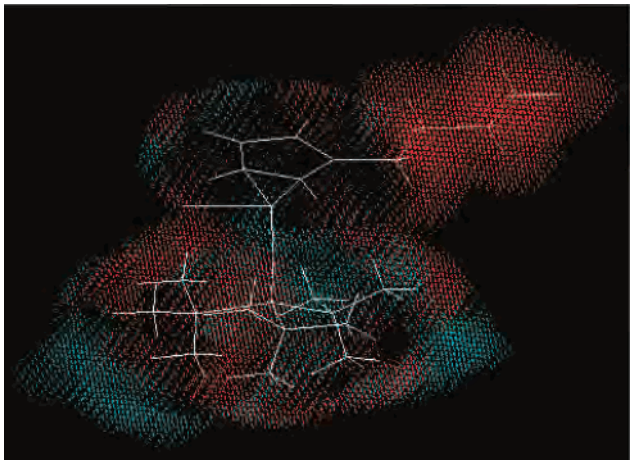


Figure 6. Illustration of the 3D-QSAR steric field constructed for Model 5 of Table 2. Red indicates areas where increased steric bulk correlates to increased activity; the cyan areas indicate areas where increased steric bulk leads to decreased activity. The structure of catalyst **28** is shown, with the ^{n}Bu group in the upper right.

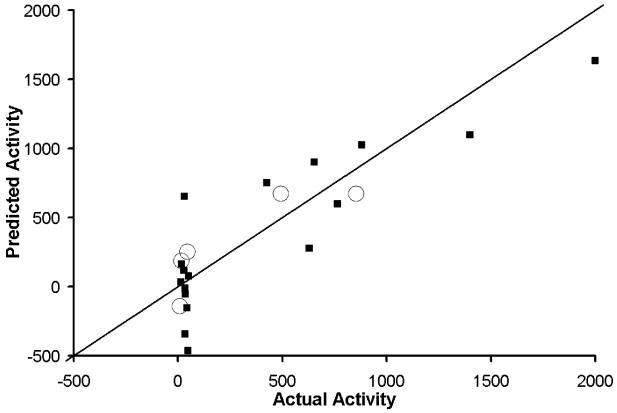


Figure 7. Actual activities for the catalysts in Table 1 plotted again their activities predicted by 3D-QSAR Model 5 of Table 3. The line is $x = y$, indicating perfect correlation. Catalysts in the training set are indicated with filled squares; the members of the test set are shown by open circles.

such a severe impairment on putative catalytic activity. However, despite these noted doubts about the numerical accuracy of the results, the 2D-QSAR results may prove to be qualitatively valid, and thus, two further observations will be offered. First, as opposed to the observation with 3D-QSAR, the newly proposed R-type ligands are all better than ^{t}Bu when Cp′ = Cp. Second, all newly proposed catalysts, even those with only an R-type substitution, have predicted activities greater than the most active original catalyst, **28** (predicted activity = 2024). This result again differs from the findings of 3D-QSAR. In order to definitively determine which finding is correct, further experimental results are required; however, absent such work, some insight may be possible by considering the strengths and drawbacks of each technique.

IV. Discussion

Applicability of 3D-QSAR to Rational Organometallic Catalyst Design. Perhaps the most disappointing result obtained with the 3D-QSAR approach is that no newly proposed catalyst has a predicted activity higher than that predicted for **28**. Figure 6 provides a graphical representation of a typical 3D-QSAR model and serves to illustrate why this is so. In addition to the patchwork of red areas (where increased bulk leads to increased activity) and cyan areas (where increased bulk leads to decreased activity) observed for most of the molecule, there is also a solidly red area in the upper right, corresponding to the ^{n}Bu group of **28**. This catalyst is the only species that contains steric bulk in this area, and given that its activity (2000) is so much higher than the second most active catalyst, **22** (1400), this substituent tends to dominate the steric field. If a large Cp′-type ligand, such as those given in Charts 2 and 3, is proposed as a new catalytically active ligand, proper alignment leads not only to population of the red, highly active ^{n}Bu portion of the 3D map but also to population of cyan portions of the map around Cp′ in addition to black portions where 3D-QSAR simply does not measure any correlation between steric bulk and activity. Thus, in essence, the failure to predict any catalysts superior to **28** is nothing more than sampling error. If the experimental data on which the 3D-QSAR model is trained were expanded to include additional ligands similar in size to the $C_5H_4(^{n}Bu)$ group of **28**, this error could be rectified.

However, the 3D-QSAR technique is still of use absent supplemental experimental data. As can be seen in Table 1, 14 catalysts possess low (i.e., <50) experimentally determined activities. There is then a gap until **24**, with an experimentally measured activity of 425, which can be defined as the beginning of the set of catalysts with medium activity. In Figure 7, the actual activities of the catalysts in Table 1 are plotted against their activities predicted with fluxional 3D-QSAR Model 5 (Table 5). Of the 14 experimentally low activity catalysts, Figure 7 indicates that eight are predicted by 3D-QSAR as possessing an activity of <50, with another five predicted to be <425, the experimental threshold for the medium activity catalysts. Indeed, only one low activity catalyst, **21**, is identified as a catalyst of medium activity. It therefore appears that the 3D-QSAR technique described in this work, at least for the Ti−N═P system, can act as an effective virtual screen, capable of identifying, before investment of experimental resources, low activity catalysts not worth synthesizing.

Applicability of 2D-QSAR to Rational Organometallic Catalyst Design. The 2D-QSAR predicted activities of Table 6, which differ from both the experimentally measured activities of Table 1 and the 3D-QSAR predicted activities of Table 5 by orders of magnitude, may seem to be grounds for dismissal of the 2D-QSAR approach as useless in predicting organometallic catalyst activities. However, the current results seem to yield useful trends, if not useful quantitative results, as shown in Table 7, where the 10 most active catalysts, as predicted by the best 2D-QSAR model (Model 4 of Table 4) are described. Also given are the rankings for these new catalysts, out of a total of 69, predicted with the other 2D-QSAR models. As can be seen, there is a great deal of consensus among the six 2D-QSAR models, despite the elimination of different descriptors for each model using the GA, different identities of the LVs

created with PLSR, and different weightings of the descriptors within each LV. In other words, it appears that the procedure in this paper may be able to successfully identify which chemical properties/descriptors correlate most with catalytic activity, even if the numerically extrapolated formulas used to predict activities of new catalysts are flawed.

Evaluation of the Database Mining Techniques. The stated reason for using database mining in this work is to identify a small number of potentially high activity ligands within a far larger set of ligands. Predicted activities for new catalysts based on these newly offered ligands are mixed: 2D-QSAR indicates an unmitigated success; 3D-QSAR indicates a failure to identify useful R-type ligands, although new Cp′-type ligands seem promising. However, the Cp′-type ligands indicated in Charts 2 and 3 all possess a common feature that represents a crucial drawback: at least one, and often many, basic donor sites that will tend to preferentially coordinate to the acidic Ti site, preventing the metal−aryl interaction at work in the Cp′−Ti−N═P framework. Even if these sites could somehow be sterically blocked from attacking Ti, these pending groups may be easily protonated, leading to complications in synthesis and purification. Although one could propose modifying the ligands in Charts 2 and 3 to remove these problematic substituents, such an approach defeats the purpose of database mining as an automated means to identify promising ligands. Given that "synthesizability" is not a solved problem even in the far more mature realm of drug design, it seems perhaps overly optimistic to expect that database mining will, in a black box fashion, result in ligands immediately ready to use in catalyst design; chemical intuition is still a necessary ingredient for identifying promising avenues of catalyst refinement. It may be promising to consider mining databases that contain ligands known to be amenable to organometallic synthesis, such as ones derived from the CSD, rather than, as was the case in the current work, databases whose compatibility with organometallic synthesis is lacking or unknown.

V. Conclusions

In the case study of this work, the experimentally characterized activities of catalysts based on a synthetically flexible Ti−N═P framework were rationalized using two techniques common to drug discovery: 3D- and 2D-QSAR, as implemented in freely available and easy to use software applied with a black box approach. In the former technique, molecular shape is taken as determinative of catalyst activity; in the latter, more general molecular properties in general are considered. For the specific system, 3D-QSAR results were found to depend on proper treatment of the highly mobile Cp′-type ligand. However, once this condition was accounted for, models were constructed with reasonable statistical measures of significance. The black box approach to 2D-QSAR necessitated the replacement of Ti with C; difficulties inherent to the technique, such as the choice of how many latent variables to include in the model, also required careful consideration. Of the final models constructed, only one showed useful levels of statistical significance. Two database mining procedures were considered to identify ligands that may lead to highly active catalysts when used as R- or Cp′-type ligands. For the former site, ligands identified thusly were trivial; for the latter, ligands seemed to be of dubious experimental significance. Both of these shortcomings can likely be addressed through a more intelligent choice of ligand database to be mined, as well as judicious application of chemical intuition to select synthetic avenues likely to prove fruitful.

Even without such ingredients, however, new catalysts based on mined ligands were constructed in silico to evaluate the applicability of the QSAR approaches for predicting activities. For 3D-QSAR, no difficulties in the black box approach were identified beyond the necessity to account for Cp ring whizzing. Due to an unbalanced steric field, one spatial area dominated constructed models, and thus, no catalysts were identified with predicted activities better than the best already known catalyst. However, this finding is not inherent to the approach, and evidence indicates that 3D-QSAR should prove useful, at the very least, in identifying low activity catalysts. For 2D-QSAR, the black box approach proved especially limiting, as not only was the method unable to accommodate Ti but also the molecular properties upon which the models were based were developed to apply to organic molecules and thus might be misleading or uninformative if a metal atom is introduced. It is expected that custom-designed descriptors, such as have been used elsewhere,[21,50] should help obviate this difficulty. Given these caveats, it is perhaps unsurprising that the different 2D-QSAR models varied tremendously in their predicted activities and that these activities were orders of magnitude away from the 3D-QSAR predictions. However, qualitative comparisons give hints that the 2D-QSAR technique may be capable of successfully identifying molecular properties influential on catalyst activity, even if such activities cannot be reliably calculated.

All in all, the computational techniques used in this study, developed to study molecules without metal atoms, are surprisingly adept at dealing with organometallic compounds where the metal atom dominates. If the limitations identified in this work—specifically a need for a balanced exploration of 3D space for 3D-QSAR model generation, the need to utilize descriptors compatible with organometallic systems for 2D-QSAR model construction, and the need to search databases of more relevance to organometallic systems—can be addressed, which should be possible in future work, computational drug discovery tools should assume an important role in identifying promising ligands usable in catalysts with new and beneficial properties.

Acknowledgment. Research sponsored by Division of Materials Science, U.S. Department of Energy under Contract No. DEAC05-00OR22725 with UT-Battelle, LLC at Oak Ridge National Laboratory. Some of this work was supported

(50) Karelson, M. In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., Eds.; Dekker Inc.: New York, 2003; p 641.

by the Center for Nanophase Materials Sciences (CNMS), sponsored by the Division of Scientific User Facilities, U.S. Department of Energy and by the Division of Materials Sciences, The extensive computations were performed using the resources of the National Center for Computational Sciences at ORNL.

Supporting Information Available: Definitions of statistical measures (r^2, q^2-CV, q^2-Test) of constructed models, possible alignments of the catalysts in Table 1, and the construction of a scramble plot of Model 4 of Table 4. This material is available free of charge via the Internet at http://pubs.acs.org.

IC700670S