Lit. Sem. 2010/6/30
Tatsuhiko Yoshino (M2)

QSAR in Catalysis ~ In Silico Catalyst Design ~

Contents

1. Introduction : What is QSAR ?

1.1 Concept

QSAR = Quantitative Structure Activity Relationships

Method to predict "activity" of target compound "quantitatively" from caluculated parameters.
Major concept in drug design. (Activity = IC_{50} etc..)

QSAR = Quantitative Structure Asymmetry Relationships
QSSR = Quantitative Structure Selectivity Relationships

1.2 Descriptor

Descriptor : Variables (parameteres) which describe the feature of molecules.
Descriptors should be obtained by experiment or calculation.

Examples : Melting point, Log P, Dipole-moment, Bond-length, Dihedral-angle, number of functional group...

QSAR model correlates Descriptor and Activity in quantitative manner.

1.3 Training/Prediction

Fitting from Experimental Results (Training)

	D_1	D_2	D_3	IC_{50}
Compound 1	100	1	0.15	30
Compound 2	50	0	0.60	3
Compound 3	75	2	1.00	0.5

Linear Correlation Model

$\log (1/IC_{50}) = c_0 + c_1D_1 + c_2D_2 + c_3D_3 + ...$

⇓

Coefficients (c_m) can be obtained by mathematical method.
(Multi-regression by least square, PLS analysis and so on)

Validation

Cross validation by LOO (leave one out) or LSO (leave several out) is usually employed for validation.

LOO (Leave One Out)

Correlation model is re-calculated with traning set where one compound is exclueded.
Then, activity of excluded compound is predicted using the new model and compared with real value.

LSO (Leave Sereral Out)

Almost same as LOO. Several compounds were excluded in this case.

Prediction Activity of other compounds can be calculated using model equation.

1.4 What is Important ?

The most important point is to select the appropriate descriptors to describe molecular structre.
For drug discovery, CoMFA is one of the most general method.

Application of QSAR methodology to predict enantio-selectivity is main topic in this seminar.

1.5 Advantages and Disadvantages

Compared with *Ab initio* Transition State Calculation

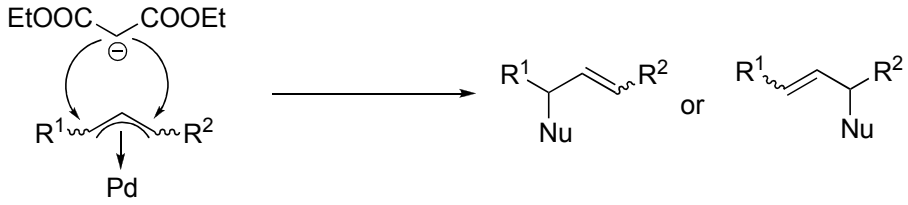
Advantages : Rapid calculation for prediction.
Easy to understand what is important for selectivity.

Disadvantages : There is no theoretical, chemical guarantee. It is only statistical result.

2. Pioneer Work : Predicting Model for Tsuji-Trost Allylation

Norrby, P-O. *et. al.* "Steric Influences on the Selectivity in Palladium-Catalyzed Allylation" *Organometallics*, **1997**, *16*, 3015.

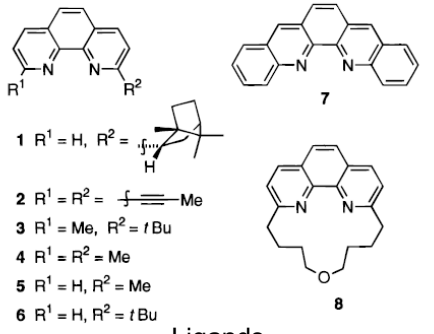
2.1 Regio-Selectivity in Tsuji-Trost Allylation



2.2 Reaction Conditions

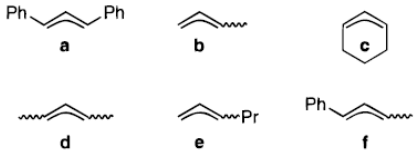
Sodium diethyl malonate + E-allylic acetate in DMF

Chart 1. Substituted Phenanthrolines Used as Ligands in Palladium-Assisted Allylation



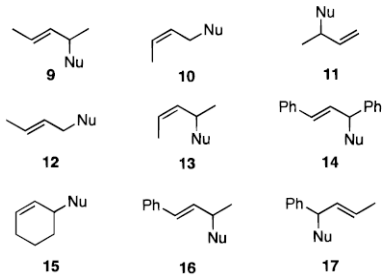
Ligands

Chart 2 η^3-Allyl Moieties Considered in This Work



Allyls

Chart 3. Isomeric Products Obtained from Allylic Substitution



Products

Catalytic and stoichiometric (using isolated π-allyls) readtions are performed.

E,*Z*-Isomerization
In catalytic reaction : fast (Boltzmann distribution by calculation)
stoichiometric reaction : slow

2.3 QSAR Model with Molecular Mechanics (MM2)

Final Selected Discriptors of Steric Feature of π-allyl Pd

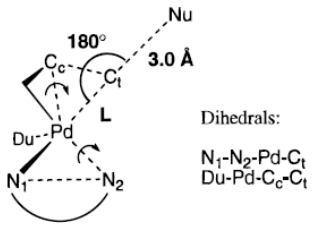


Figure 2. Descriptors in the final model. The "Du" pseudoatom is needed in the molecular mechanics description of the complex. The Pd−Du vector is approximately perpendicular to the average coordination plane.

"Selectivity" was conveted to Gibbs energy.
(Reaction Rate = K*exp(-E_a/RT))

Linear model

$$\Delta G^* = c_0 + \sum_n c_n x_n$$

Discriptors
1) Breaking Pd-C bond length
2) Dihedrals N_1-N_2-Pd-C_t
3) Dihedrals Du-Pd-C_c-C_t
4) Steric Interaction with Nu

(Structures of π-allyl Pd were generated by MM2 method)

Steric Interaction with Nu : Set **Ar probe atom** to Nu position (figure) and calculate the increased energy

Regression was performed with Levenberg-Marquardt algorithm (normal multi-regression)

2.4 Results and Discussion

Table 1. Experimental and Calculated Isomeric Ratios of the Products in Palladium-Catalyzed Allylic Substitution

			product ratio		ΔΔG*/kJ mol^{-1}	
entry	complex	products	exptl[a]	calcd[b]	exptl[a]	calcd[b]
1	**1a**	(*S*/*R*)-**14**	2.23[c,d]	1.25	1.98	0.56
2	**1c**	(*R*/*S*)-**15**	3.50[c]	2.26	3.10	2.02
3	**1d**	(*S*/*R*)-**9**	1.27[c,d]	1.88	0.60	1.56
4	**1f**	**16/17**	2.30[e]	2.02	2.06	1.74
5	anti-**2b**	**11/10**	1.86[f]	1.32	1.53	0.70
6	**3b**	**10/11**	1.00[f,g]	1.02	0.00	−0.05
7	anti-**4b**	**11/10**	1.50[f,h]	0.83	1.00	−0.47
8	syn-**4b**	**12/11**	99.0[h]	98.9	11.4	11.4
9	**4d**	**9/13**	49.0	36.6	9.64	8.92
10	**4f**	**16/17**	10.1	8.11	5.73	5.18
11	**5f**	**16/17**	1.45[e]	1.53	0.92	1.05
12	**6f**	**16/17**	4.50[e]	11.3	3.73	6.01
13	**7b**	**10/11**	2.33[f,g]	0.84	2.10	−0.42
14	**8b**	**10/11**	3.27[f,g]	4.03	2.94	3.46
			rms δΔΔG*/kJ mol^{-1}:			1.19

[a] Except where noted, results from catalytic reactions of *E*-allylic acetate with sodium diethyl methylmalonate in DMF, ref 19c. [b] *T* = 298 K. Syn−anti isomerization is assumed to be fast relative to nucleophilic attack in catalytic reactions and slow in stoichiometric reactions. For entries 5, 7, and 8, only conformers of one allyl isomer (syn or anti) were used in the calculations. [c] Result from ref 20. [d] Absolute configuration was not assigned. [e] This work. [f] Product **11** (internal attack) was assumed to result from attack on anti complex, cf. entry 8. [g] Experimental value for the hexenyl system, allyl **e**. [h] Stoichiometric reaction.

Cross Validation Value : LOO Q^2 = 0.86
LSO Q^2 = 0.87

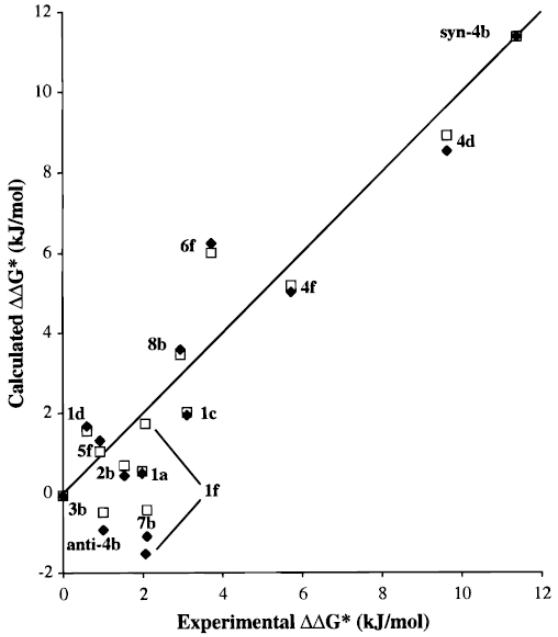


Figure 1. Correlation between calculated and experimental activation energy differences (kJ mol^{-1}) for product isomers. The values are labeled corresponding to the intermediate complexes (Charts 1 and 2, Table 1). Data points are indicated by (□) for calculations by the final model and (◆) for a predicted data point that was left out of the fitting procedure (LOO validation).

·The most important factor (descriptorw that coefficient has largest absolute value) is Pd-C bond.

·About result of cross validation of **1f** which has unsymmetrical allyls and chiral ligand

1) Error occured because the fact that crossover between enantiometric path cannot take place.
2) Error occured because asymmetrical electronic effect was neglected.

⟹ Modification for these problem did not give better models.

⟹ Just a possible Error in MM2 system ?

3. CoMFA : Analysis of Asymmetric Diels-Alder Reaction

Lipkowitz, K. B. *et. al.*
"Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels-Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands"
J. Org. Chem. **2003**, *68*, 4648.

3.1 Method

Comparative Molecular Field Analysis (= CoMFA) is now widely used method for drug design.
This method was first reported in 1988 in *JACS*.

Ref) Cramer, R. D. III *et. al.* "Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins" *J. Am. Chem. Soc.* **1988**, *110*, 5959.

Key for CoMFA

1) All the analyzed compounds are set in grid space in appropriate manner.
Then **interaction energy between "probe"** and each coumpound at all grid points.
These energy values are used as descriptors.

2) Generated huge number of descriptors are analyzed by **"PLS-regression"** technique.

Grid-based Descriptor

Probe : sp^3 C^+ atom is often utilized. (Other probes can be also utilized.)
Interaction energy is calculated as sum of **van der Waals energy** and **Coulombic energy**.

van der Waals Energy : Tripos Force Field ($\propto$**1/r^6 – 1/r^{12}**)
Coulombic energy : $\propto$**1/r**, Atomic charge by *Gasteiger-Marsili method (from **orbital electronegativity**)

Other calculation methods can be used.

* *Tetrahedron* **1980**, *36*, 3219.

Grid Point

- Enough range to cover all atoms of target compounds
- Grid points with too high steric energy is cut-off.
- Grid points with too small standard deviation is eliminated.
- Grid space is usually 1.0-2.0 Å.

PLS-Regression

PLS = Partial Least Square is a regression method suitable for models with...

- No clear reasonable relation of variables and property.
- Large number of variables (descriptors), often larger than number of samples.

Overfitting is a big problem in such a case, if normal multi-regression is employed.

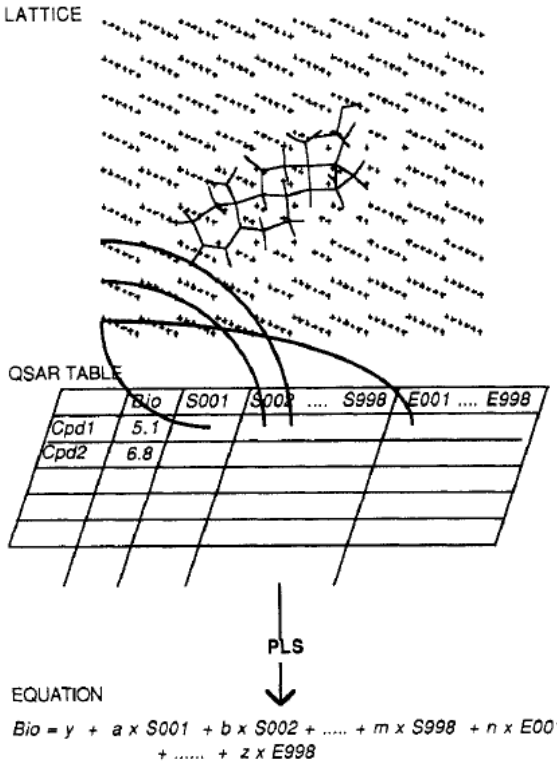


Figure 1. The process of comparative molecular field analysis (CoMFA).

PLS-Regression : m variables (descriptors), n samples, 1 output models

$$y_i = c_1 t_{i1} + c_2 t_{i2} + \cdots + c_r t_{ir} + e_i = \sum_{k=1}^{r} c_k t_{ik} + e_i$$

$$t_{i1} = w_{11} x_{i1} + w_{12} x_{i2} + \cdots + w_{1m} x_{im} = \sum_{j=1}^{m} w_{1j} x_{ij}$$

$$t_{i2} = w_{21} x_{i1} + w_{22} x_{i2} + \cdots + w_{2m} x_{im} = \sum_{j=1}^{m} w_{2j} x_{ij}$$

$$\vdots$$

$$t_{ir} = w_{r1} x_{i1} + w_{r2} x_{i2} + \cdots + w_{rm} x_{im} = \sum_{j=1}^{m} w_{rj} x_{ij}$$

Ref) http://cse.naro.affrc.go.jp/iwatah/index_j.html
PLS回帰入門

y_i : output
x_{ij} : variables j of sample i
t_{ik} : latent variables (LC of x)

w : weight coefficients for X to T
c : coefficients

Calculation of w

1) For t_{i1}, $\{w_{1j}\}$ is obtained to maximize covariance (colinearlity) of $\{y_i\}$ and $\{t_{i1}\}$.

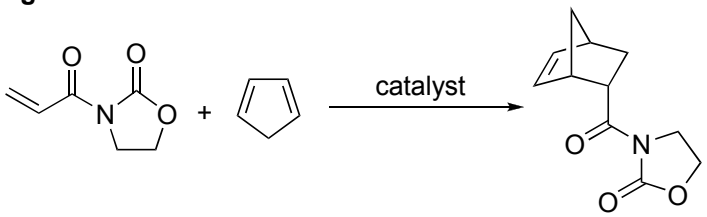
2) For t_{i2}, $\{w_{2j}\}$ is obtained to maximize covariance (colinearlity) of $\{y_i - c_1 t_{i1}\}$ and $\{t_{i2}\}$.

3) For t_{i3}, $\{w_{3j}\}$ is obtained to maximize covariance (colinearlity) of $\{y_i - c_1 t_{i1} - c_2 t_{i2}\}$ and $\{t_{i3}\}$.

4) This procedure is repeated to reach r.

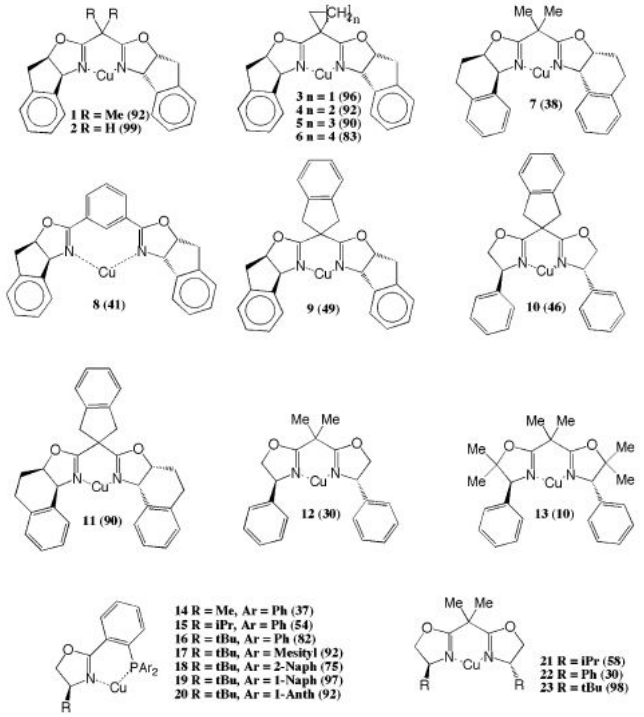
3.2 Application to Asymmetric Catalysis

Taget Reaction and Reaction Conditions



Experimental results were extrated from repoted papers.
The most optimized reaction conditions were used.
Differeces in temperature, solvent etc were not considered.

Catalysts



Counter anions are not described,
but considered in structure optimizing calculation.

Catalyst Structure

Initial Structures : CSD or Built in Spartan

⇓

Optimized by PM3tm method (semiempirical MO)

Alignment

Least-square fitting of oxazoline rings.

Validation

1) Internal Cross Validation by LOO.
In some cases, only internal validation is not enough.
Ref) Golbraikh, A; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.

2) External Validation
At ramdom, **10**,**15**,**18**,**22** were excluded and used as external prediction set (LSO).

Best model

Field : both (Steric and Electronic)
Energy cut-off : S/E =30/20 kcal/mol
dielectric functions $1/r^2$
Probe : C^+sp^3
Latent variables : 6

1) All catalysts model : $r^2_{cv} = 0.833$
2) LSO catalysts model : $r^2_{cv} = 0.785$, external $r^2 = 0.94$
Golbraikh Tropsha Criteria : fullfilled

Visualized Results (STDEV*COEFF contour plot)

All Catalysts

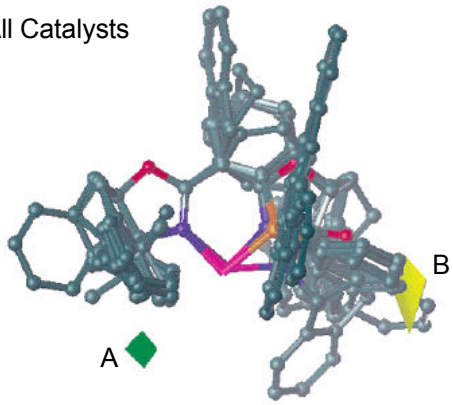


FIGURE 3. CoMFA steric STDEV*COEFF contour plot. Shown inside the field is the aligned set of 23 chiral catalysts with hydrogen atoms removed for clarity. Placement of bulky groups near the green region (contoured at contribution level 93) and/or removal of steric bulk near the yellow region (contoured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

Good Catalyst

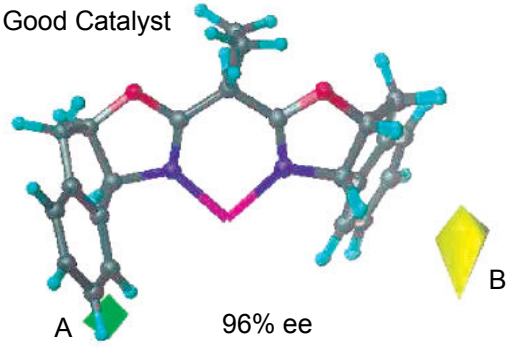


FIGURE 4. CoMFA steric STDEV*COEFF contour plot. Shown inside the field is the highly efficient catalyst **3** (ee 96%). It is to be noted that significant steric bulk lies in the green region while the yellow region is devoid of steric bulk confirming the model.

Bad Catalyst

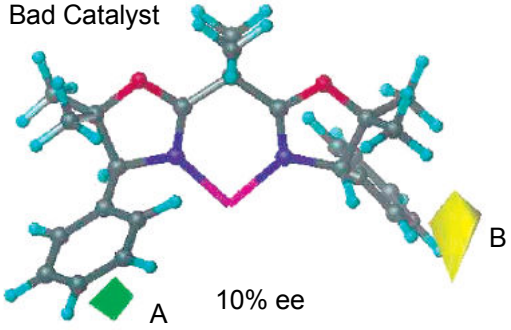


FIGURE 5. CoMFA steric STDEV*COEFF contour plot. Shown inside the field is the inefficient catalyst **13** (ee 10%). It is to be noted that while significant steric bulk lies in the green region the yellow region has too much steric bulk that, in turn, reduces the effectiveness of this catalyst.

Contribution of Each Factor to Selectivity

Steric : 60–70%
Electronic : 30–40% $\Longrightarrow$ Steric factor is more important.

For Higher Selectivity

A : Steric hindrance should be **increased.**
B : Steric hindrance should be **decreased.**

4. QM-QSAR : Works of Prof. Marisa C. Kozlowski

4.1 Method

Dixon, S. L.; Merz, K. M. Jr. *et. al.* "QMQSAR: Utilization of a Semiempirical Probe Potential in a Field-Based QSAR Method" *J. Comp. Chem.* **2004**, *26*, 23.

Compared with CoMFA...

Similar Point : Energy values at grid points are calculated and used as descriptors.

Different Points : Energy is calculated as Probe Interaction Energies (PIE) by **quantum mechanical method.** Regression is performed using **n-variables regressions by simulated-annealing**.

Probe Interaction Energies (PIE)

Probe : **a positively charged carbon 2s electron**

$$\mathrm{PIE}(g_i) = -\langle s_i s_i | V(L) \rangle = \int_{\mathbf{r}_1} \chi^*_{s_i}(\mathbf{r}_1)\chi_{s_i}(\mathbf{r}_1)\left\{ \sum_{\alpha \in M} \left[\frac{Z_\alpha}{|\mathbf{r}_1 - \mathbf{r}_\alpha|} \right. \right.$$

Potential from nucleus

$$\left. \left. - \sum_{\mu \in \alpha} \sum_{\mu' \in \alpha} P_{\mu\mu'} \int_{\mathbf{r}_2} \frac{\chi^*_\mu(\mathbf{r}_2)\chi_{\mu'}(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_2 \right] \right\} d\mathbf{r}_1 \quad (3) \qquad P_{\mu\mu'} = 2 \sum_{k=1}^{N_{occ}} c_{\mu k} c_{\mu' k}$$

Potential from electrons

M : Considered atoms
α : nucleus
χ_{si} : wave functions of probe
χ_μ : wave functions of basis in α
c : coeffiecents of LCAO

All calculations are performed PM3 (semiempirical MO) method.

Regression

From several thousands of descriptors, n (2,3,4,5...) descriptors which give good fitting are selected.

How to select optimal descpriptors

Simulated Annealing Ref) *Science* **1983**, *220*, 671.
Review) *Eur. J. Oper. Re.* **1990**, *46*, 271.

Solutions for optimization problems in NP-hard class.
(impossible to solve in polynomial-time by deterministic algorithmm, if N≠NP)
Solution time is polinomial, but there is *no guarantee to always give the right answer.*
Mimic of annealing process.

1) T (Temperature) and initial state is set.
2) "State" is changed to another neighbour state "stochastically" as following.
   If next state is better than now, state transition occured.
   If next state is worse and temperature is enough high, state transition occured. (A)
   These operations are repeated.
3) Temperature is decreased.
4) Repeat 2)-3)

By process **(A)**, probability of wrong answer which is "local minimum" decreased. (Think of start from P.)

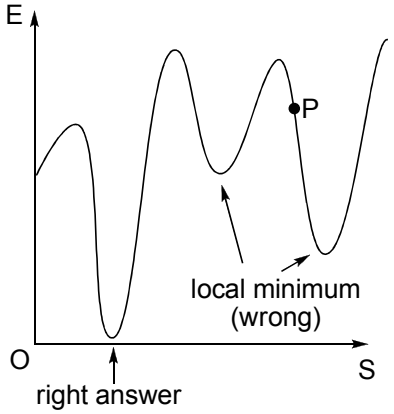


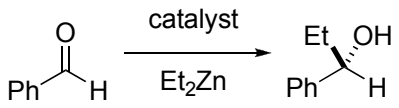
Table 2
Simulated Annealing algorithm in pseudo-code

```
Select an initial state i ∈ S;
Select an initial temperature T > 0;
Set temperature change counter t = 0;
Repeat
Set repetition counter n = 0;
  Repeat
    Generate state j, a neighbour of i;
    Calculate δ = f(j) − f(i);
    If δ < 0 then i := j
      else if random(0, 1) < exp(−δ/T) then i := j; (A)
    n := n + 1;
  until n = N(t);
t := t + 1;
T := T(t);
until stopping criterion true.
```

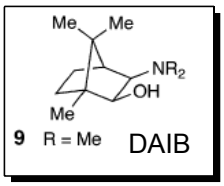
4.2 QM-QSAR Approach for Predicting the Selectivity of Asymmetric Alkylation

Kozlowski, M. C. et. al. "Quantum Mechanical Models Correlating Structure with Selectivity: Predicting the Enantioselectivity of β-Amino Alcohol Catalysts in Aldehydee Alkylation" *J. Am. Chem. Soc.* **2003**, *125*, 6614.

Target Reaction

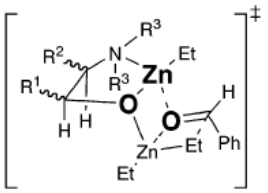


99% ee



Noyori, R. *et. al. J. Am. Chem. Soc.* **1986**, *108*, 6071.

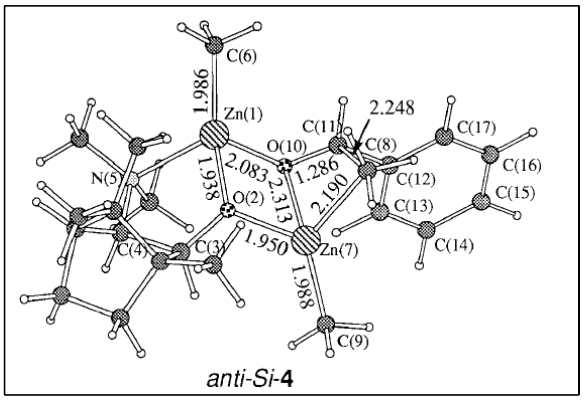
Transition State



anti S

Among possible 4 TS (*syn*/*anti*, *R*, *S*), *anti S* is the most favored.

Calculated TS (using Me_2Zn) at the RHF/3-21G/Zn level.



anti-Si-**4**

Noyori, R. *et. al. Organometallics* **1999**, *18*, 128.

Catalysts Set

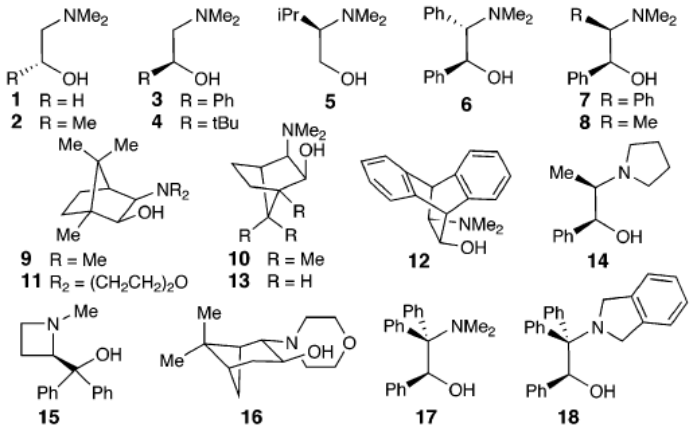


Figure 2. β-Amino alcohol catalysts.

Catalysts Structure Optimization

Optimized Ground States : planar Zn

⇓

not suitable

⇓

Transition States (PM3)

⇓

Aligned fitting Zn-O-Zn atoms

Method

·PIEs (descriptors) are correlated with **ΔG** values which can be converted to **ee**.

ΔG = RT ln K, where K is enatio metric ratio.

·2 PIEs are selected to give the best fitting model by simulated annealing and normal least-square is used.

$\Delta G = a + c_1(PIE_1) + c_2(PIE_2)$ ⟶ *Best Individual Model*

Or all acceptable PIEs pairs are weight averaged. ⟶ *Averaged Model*

Effect of Grid Space

Table 2. Statistical Summary of the QSSR Models

TS[a]/grid spacing	model	RMSE[b]	R^{2} [c]	CC[d]	predicted R^{2} [e]	N[f]
grid2/2.0 Å	best	0.81	0.23	0.50	0.32	54
	avg	1.27	-0.87	-0.29	-0.66	54
grid2/ 1.3 Å	best	0.29	0.90	0.99	0.92	174
	avg	0.34	0.86	0.98	0.88	174
grid2/ 0.7 Å	best	0.34	0.86	0.93	0.88	1077
	avg	**0.30**	**0.90**	**0.95**	**0.91**	1077
grid1/0.7 Å	best	0.49	0.72	0.95	0.75	1036
	avg	**0.29**	**0.90**	**0.96**	**0.92**	1036

⟹ 0.7 Å grid space show good convergence.

CC = Correlation Coefficients

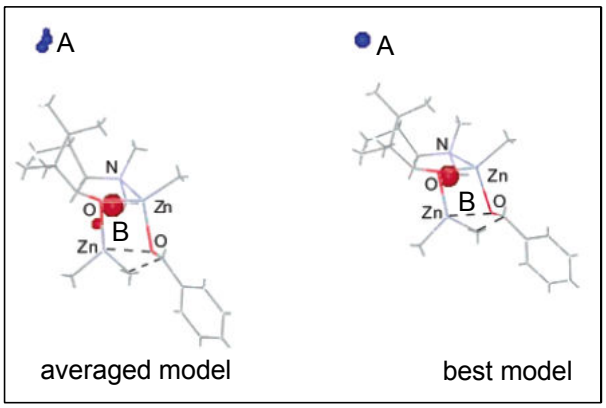
Describe how well the prediction set selectivity ***order*** is calculated.

Results

Table 1. QSSR Calculations Using Catalysts from **1**–**18**[a]

	expt.		*anti S*1 best				*anti S*1 avg	
cmpd	% ee[b]	ΔG^c	% ee_{fit}	ΔG_{fit}[d]	PIE_1[e]	PIE_2[e]	% ee_{fit}	ΔG_{fit}
				Training Set[f]				
1	0	0.00	28	0.32	25.92	4.76	26	0.30
3	59	0.76	70	0.97	21.79	3.54	68	0.92
4	93	1.85	89	1.61	20.62	4.44	90	1.63
5	49	0.60	24	0.27	26.25	4.87	34	0.39
6	66	0.88	72	1.00	21.92	3.72	70	0.96
7	73	1.04	72	1.00	24.28	5.44	76	1.10
8	81	1.26	79	1.18	23.15	5.10	81	1.26
9	98	2.56	98	2.44	18.39	5.11	98	2.47
10	95	2.04	97	2.35	18.53	4.95	97	2.38
11	98	2.56	97	2.35	19.14	5.39	97	2.37
12	96	2.17	97	2.26	21.55	6.94	96	2.18
13	94	1.94	91	1.69	20.92	4.88	93	1.83
17	94	1.94	95	2.02	21.35	6.12	93	1.89
18	97	2.33	97	2.43	20.97	6.97	96	2.23
				Prediction Set[g]				
2	3	0.03	11	0.12	27.91	5.67	5	0.06
14	86	1.43	81	1.25	21.45	4.05	76	1.10
15	98	2.56	99	3.36	15.29	5.36	99	2.82
16	63	0.83	83	1.31	23.66	5.85	75	1.09

[a] Catalyst geometries taken from *anti S* transition structures. Grid1 orientation, 0.7 Å grid spacing. [b] (*S*)-product. [c] The % ee is converted to ΔG (kcal/mol) using $\Delta G = RT \ln K$, K is ratio of the (*R*) and (*S*) enantiomers. [d] $\Delta G_{fit} = a + c_1(PIE_1) + c_2(PIE_2)$; $a = 5.48$ kcal/mol, $c_1 = -0.27$, $c_2 = 0.36$. [e] Probe interaction energies (kcal/mol) at the two grid points identified in the QSSR analysis. [f] best,avg: SD = 0.23, 0.17 kcal/mol; $R^2 = 0.93$, 0.95. [g] best,avg: RMSE = 0.49, 0.29 kcal/mol; $R^2 = 0.72$, 0.90; CC = 0.95, 0.96.



For good Selectivity.... A: more PIEs
B: less PIEs

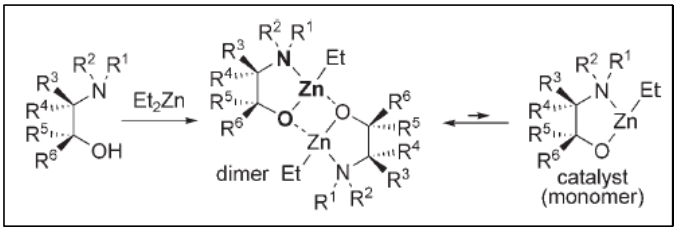
PIEs : Electron rich area : decreased
Near Nucleus : inceased

Only minutes of computing gave good models!

4.3 Further Prediction for New Catalysts

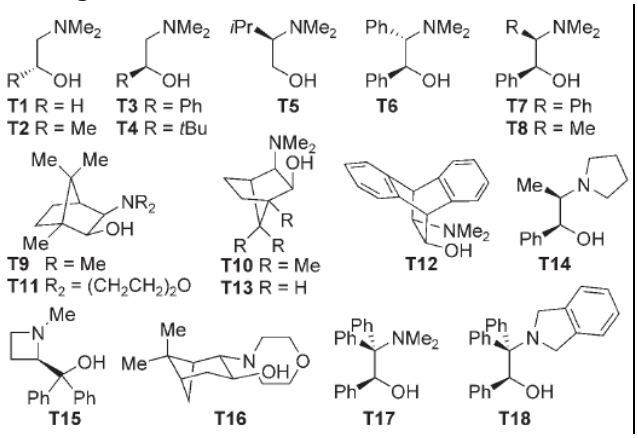
M. C. Kozlowski *et. al.* "A Priori Theoretical Prediction of Selectivity in Asymmetric Catalysis: Design of Chiral Catalysts by Using Quantum Molecular Interaction Field" *Angew. Chem., Int. Ed.* **2006**, *45*, 5502.

Method Improvement : **Ground states** of dimeric catalysts gave a good model.

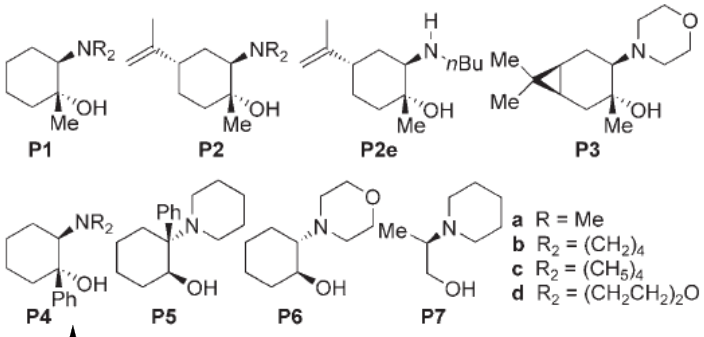


Monomer : Trigonal Zn ... Bad model
Dimer : Tetrahedral Zn ... Good model

Training Set



Prediction Set



Trans amino alcohols with tetrasubstituted chiral centers were included.

Results

Table 1: QSSR a priori predictions for catalysts derived from **P1**–**P7**.

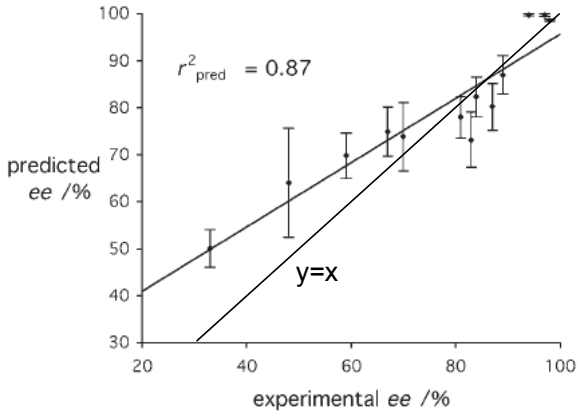
Ligand	Single run[a] ΔG[c]	Single run[a] % *ee*[d]	"Leave-two-out"[b] Mean ΔG[c]	"Leave-two-out"[b] SD ΔG[c]	"Leave-two-out"[b] % *ee*[S]	CI % *ee*[e]	Expt. % *ee*[f]
P1a	1.74	92.3	1.68	0.22	91.5	3.2	–
P1b	0.81	63.5	0.91	0.33	68.8	15.9	–
P1c	1.16	78.9	1.45	0.38	87.2	8.2	89
P1d	0.80	63.0	1.01	0.28	73.2	11.8	83
P2a	1.01	73.1	1.05	0.26	74.9	10.3	67
P2b	0.92	69.3	1.03	0.35	73.9	14.6	70
P2c	1.14	78.5	1.19	0.31	80.2	10.0	87
P2d	1.00	72.7	1.13	0.24	78.0	8.7	81
P2e	0.74	59.2	0.82	0.43	64.1	23.1	48
P3	1.11	77.4	1.26	0.29	82.3	8.4	84
P4a	2.88	99.0	2.68	0.31	98.6	0.8	–
P4b	2.10	96.0	2.01	0.38	95.2	3.2	–
P4c	3.61	99.7	3.77	0.77	99.8	0.3	97
P4d	3.67	99.8	3.78	0.75	99.8	0.2	94
P5	2.71	98.7	2.69	0.16	98.6	0.4	98
P6	0.89	67.6	0.93	0.21	69.8	9.8	59
P7	0.47	41.2	0.55	0.24	50.1	8.0	33

[a] Compounds **T1**–**T18** served as the parameterization set. [b] "Leave-two-out" cross-validating analysis using 16 compounds from **T1**–**T18** as the parameterization set (all 153 combinations). [c] Calculated ΔG in kcal mol^{-1}. [d] Generated from ΔG at 273 K; *S* enantiomer product. [e] 95% confidence interval. [f] From reactions performed at 273 K. *S* enantiomer product.

"Leave-two-out"
All 153 combinations of 16 catalyst from **T1**-**T18** gave 153 models.

SD = Standard deviation
CI = 95% confidence interval

Summary of Prediction

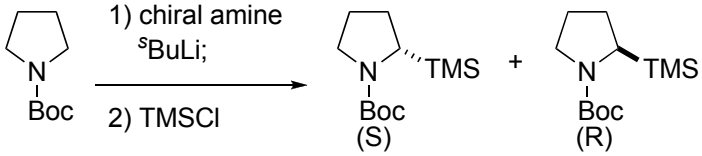


Leave-two-models gave better correlatioon.

4.4 G-QSAR Approach for Asymmetric Lithiation of *N*-Boc-pyrrolidine

Kozlowski, M. C *et. al.* "Is the A-Ring of Sparteine Essential for High Enantioselectivity in the Asymmetric Lithiation-Substitution of *N*-Boc-pyrrolidine ?" *J. Am. Chem. Soc.* **2004**, *126*, 15473.

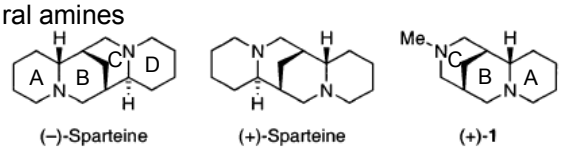
Target Reactions



(–)-sparteine : *(S)* >90% ee
(+)-1 : *(R)* >90% ee

(+)-1 is easily synthesized and works as **(+)-sparteine** surrogate.

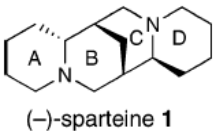
chiral amines



O'Brien, P. *et. al. J. Am. Chem. Soc.* **2002**, *124*, 11870.

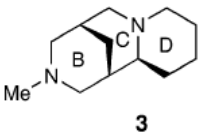
Mechanistic Investigation of **(–)-Sparteine**/sBuLi System
Wiberg. K. B.; Bailey. W. F. *J. Am. Chem. Soc.* **2001**, *123*, 8231.

Is A ring essential ?



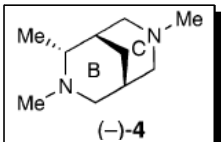
(–)-sparteine **1**

(S) 95%ee



3

(S) 21% ee



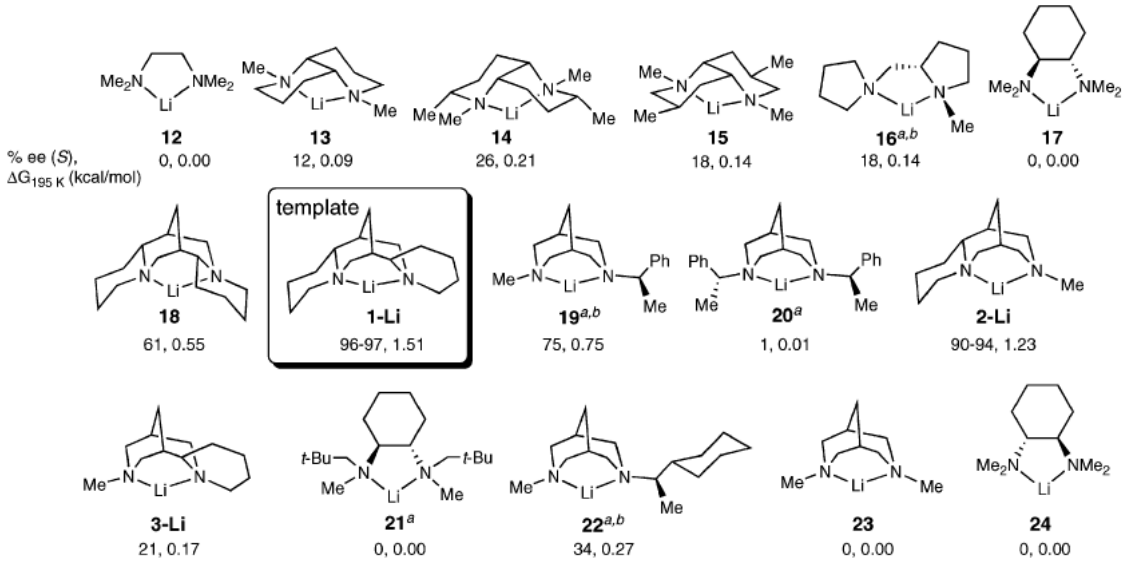
(–)-**4**

(R) 35% ee

What is important for selectivity ?

Training Set

Chart 1. Training Set Diamine Lithium Complexes and Their Enantioselectivities from Eq 1 for the Asymmetric Lithiation–Substitution QSSR[c]



[a] The lowest energy conformation of the lithium complex was employed in the QSSR analysis. [b] Two orientations were evaluated in the QSSR analysis. The illustrated orientation gave the best models. [c] ΔG values were obtained from $\Delta G = -RT \ln K$, where $K =$ er and corresponds to the differences in energy between two pathways leading to the enantiomeric products.

Results

No good correlation model was obtained by QM-QSAR with PM3 calculation.

⇓

More precise calculation of PIEs was necessary.

⇓

G-QSAR : PIEs can be calculated using...
appropriate method (HF, MP2, B3LYP)
appropriate basis set (3-21G, 6-31G*, 6-31+G**...)
with Gaussian program.

Optimized Method

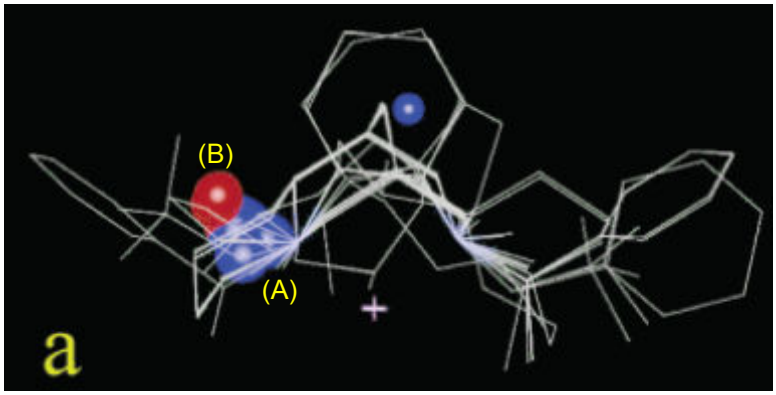
Structure Optimization : HF/3-21G* (*ab initio* MO)
PIE calculation : B3LYP/6-31G** (DFT)
2-variables model
LOO Cross Validation : $r^2_{cv} = 0.68$
Correlation Coefficient (CC) = 0.82

By optimized model, **(–)-4** was predicted to give *(R)* product in 22-25% ee (exp. *(R)* 35% ee)

Table 2. Experimental vs Predicted Enantioselectivity and ΔG Values for the Reaction in Eq 1 Using the DFT QSSR Model

	ΔG (kcal/mol)		% ee	
complex	expt[a]	pred[a,b]	expt	pred[b]
12	0	−0.032	0	4 (*R*)
13	0.093	−0.001	12	0.1 (*R*)
14	0.206	−0.199	26	25 (*R*)
15	0.141	0.118	18	15 (*S*)
16	0.141	0.279	18	34 (*S*)
17	0	0.245	0	30 (*S*)
18	0.549	0.825	61	78 (*S*)
1−Li	1.508	1.502	96−97	96 (*S*)
19	0.754	0.214	75	27 (*S*)
20	0.008	0.666	1	69 (*S*)
2−Li	1.232	1.140	90−94	90 (*S*)
3−Li	0.165	0.279	21	34 (*S*)
21	0	−0.151	0	19 (*R*)
22	0.274	0.460	34	53 (*S*)
23	0	0.149	0	19 (*S*)
24	0	−0.056	0	7 (*R*)

[a] ΔG obtained from $\Delta G = -RT \ln K$, where $K =$ er and corresponds to the differences in energy between two pathways leading to the enantiomeric products. [b] Obtained from the leave-out-one cross-validation QSSR analysis for the complexes in Chart 1.



For better selectivity...

(A) More PIEs
(B) Less PIEs

These grid points are located above/below A ring.

1) Large group below A ring : good←(A)
2) Large alkyl group above A ring : Bad←(B)
3) Ph group above A ring : good←(B)

Structure aroud A ring seemed esseintial !

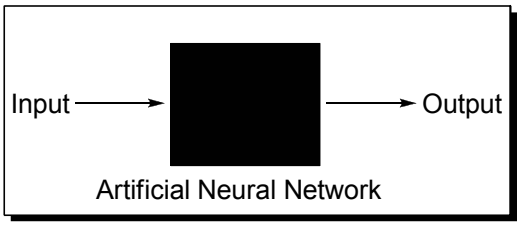
Other application example : Kozlowski, M. C. *et. al. Org. Lett.* **2006**, *8*, 1565.

5. Neural Network : Non-Linear Regression Methodology

5.1 Method : Artificial Neural Network Model

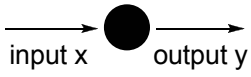
Serra, J. M. *et. al.* "Can artificial neural networks help the experimentation in catalysis?" *Catalysis Today* **2003**, *81*, 393.
「化学者のためのニューラルネットワーク入門」、1996年、ユーレ・ジュパン ヨーハン・ガスタイガー、丸善

General Concept



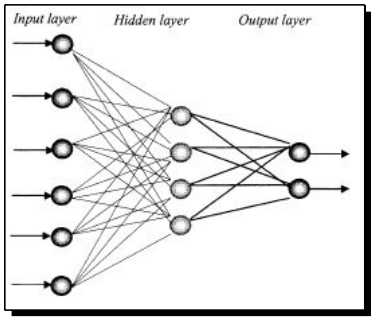
Artificial Neural Network (ANN) works as "black box",
which gives "Output" from "Input"
even if the correlation is extremely complex and unknown.
"Black box" is programed to mimic a neural network(brain).

Artificial Neuron Model



Linear model : $y = x$
Non-linear model : $y = 1/(1+e^{-x})$ These are called "activation function".
$y = (e^{x}–e^{-x})/(e^{x}+e^{-x})$

Network Model : Multi-Layer Perceptrons



Weighted sums of previous layers' outputs are used as next inputs.

next input of neuron (i)

$$x_i = w_{i1}y_1^{prev} + w_{i2}y_2^{prev} + w_{i3}y_3^{prev} + \ldots + w_{ij}y_j^{prev}$$

This calculations are performed for all next neourons.

input for next neurons : $\mathbf{X} = {}^{t}\{x_1, x_2, \ldots, x_i\}$
previous out put : $\mathbf{Y^{prev}} = {}^{t}\{y_1^{prev}, y_2^{prev}, \ldots, y_j^{prev}\}$
weight matrix $\mathbf{W} = \{w_{ij}\}$

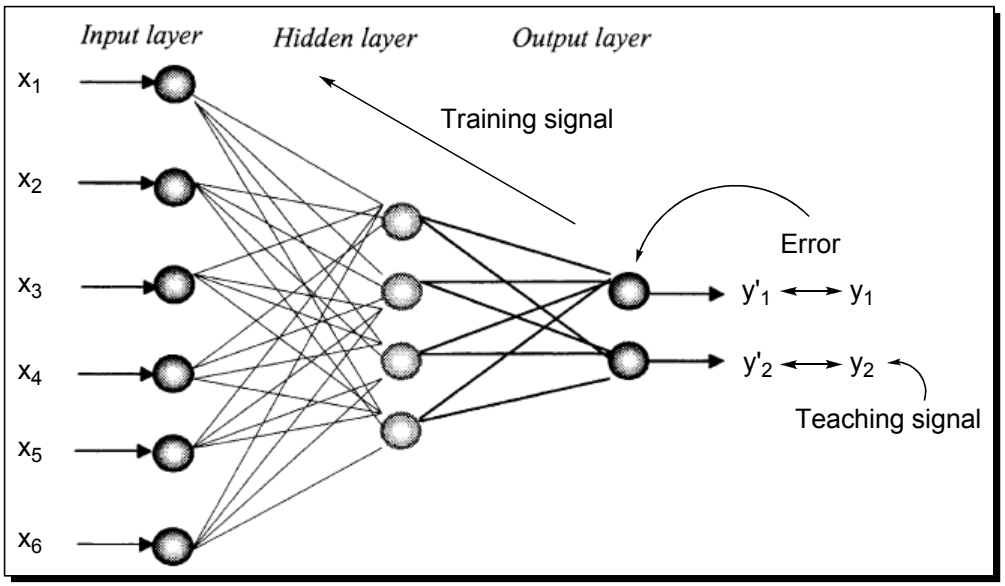
$\Rightarrow \mathbf{X = WY^{prev}}$

next output $\mathbf{Y} = g(\mathbf{X})$ (g : activation function)

Weight matric **W** should be optimized to give good correlation.
(N-1) Matrixes exist in N layers model.

Backpropagation - Concept

When a teaching signal (**X**, **Y**) is given, weight matrixes **W** are modified to minimize errors.



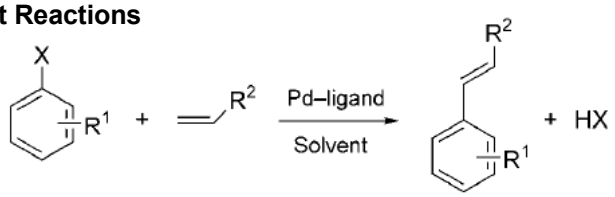
Advantage ·Applicable to many problems where theoretical anaysis or linear-regression is difficult.

Disadvantage ·It it impossible to obtain theoretical or qualitative information from the results.

5.2 QSAR Investigation of Heck Reactions

Farrusseing, D.; Rothenberg G. *et. al.* "Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions" *Adv. Synth. Catal.* **2004**, *346*, 1844.

Target Reactions



R^1 = H, OH, CHO, Me, OAc, OBz, NH_2, OMe, NHCOMe, NO_2, CN, COMe, CO_2Me, F, $N(Me)_2$, CF_3

R^2 = CO_2Bu, CO_2Me, Ph, CO_2H, $(CO_2)Et$, $CON(Et)_2$, CN

Scheme 1. The general Heck reaction described by the data-set. Ligands used are monophosphines and monophosphites; solvents are DMF, THF, DMA, dioxane, Et_3N, PhMe, NMP, MeCN, EtCN, PrCN, HMPT, and 1,2-DCE.

·412 Reactions were collected to analyze from reported papers with various conditions.

·Activity : log(TON) and log(TOF)

Descriptors

Initial Set (76 descriptors)
Steric descriptors : MW, Surface, Volume, Tolman's cone angle, Solid angle and related parameters etc
Electronic descriptors : Hammett constant, HOMO, LUMO, GAP, Dipole moment, Chages on ligating atoms etc
Others : Pd loading, Pd precursor, reaction time, Temperature

Selected Descriptors Set (reduced by Relief Algorithm and Principal Component Analysis)

For TON (17 descriptors)

R_1(Halide) : HOMO, LOMO, GAP, S_{occ}
R_2(Olefin) : LUMO, GAP, dipole, A
Ligand : q, HOMO, LUMO, GAP, S_{occ}
Solvent : q
Others : Temp, Pd loading, Cat. precursor

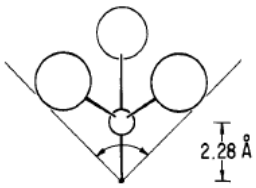
For TOF (20 descriptors)

R_1(Halide) : $Hammet_{p(+)}$, $Hammet_{p(-)}$, V
R_2(Olefin) : HOMO, LUMO, V, S(ethylene)/S, O, dipole
Ligand : q_2, HOMO, LUMO, S_{occ}, A, R_{max}
Solvent : O_{max}
Others : Temp, Pd loading, Cat. precursor, Time

q : chage on ligating atom

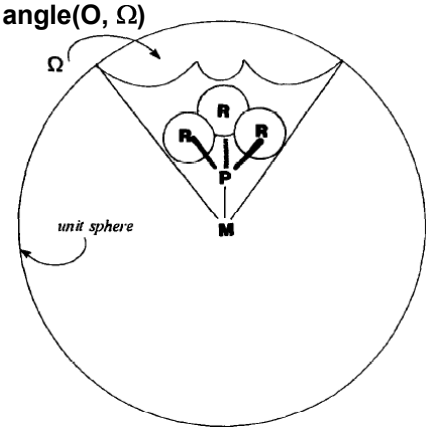
S_{occ} : percentage of sphere occupation

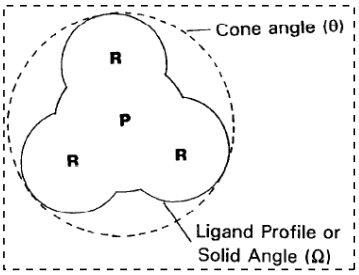
Tolman's Cone Angle (Θ, T)



Tolman, C. A. *Chem. Rev.* **1977**, *77*, 313.

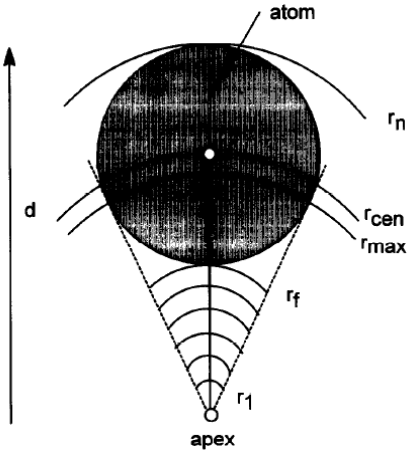
Solid angle(O, Ω)

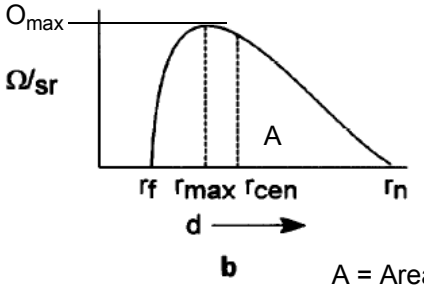




Solid angle is reflected by the shape of ligand.

Solid angles for radial profile (O_{max}, A, R_{max})



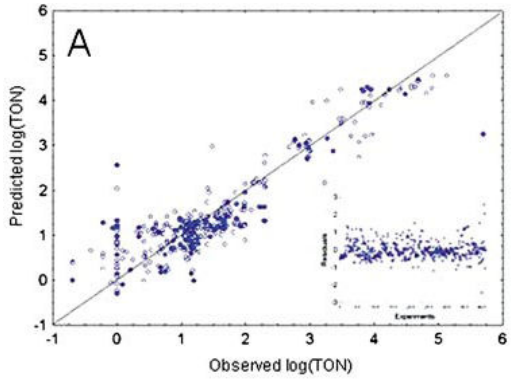


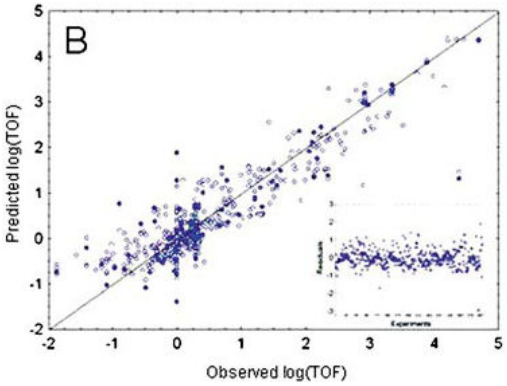
A = Area under the curve

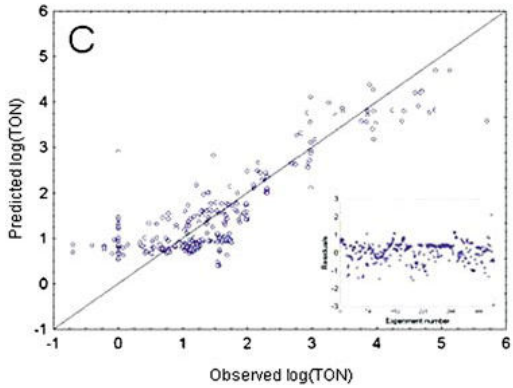
White, D. *et. al. J. Organomet. Chem.* **1994**, *478*, 205.

Generated ANNs For TON : 11 nodes and 3 nodes
For TOF : 15 nodes and 10 nodes in the 1st and 2nd hidden layers.

Reluts







A : TON by ANNs
B : TOF by ANNs
C : TON by linear regression model

Inset : residuals

ANNs > Linear regression model ?

Classification Problem

Table 1. Confusion matrix results for classification analyses of TON and TOF values.

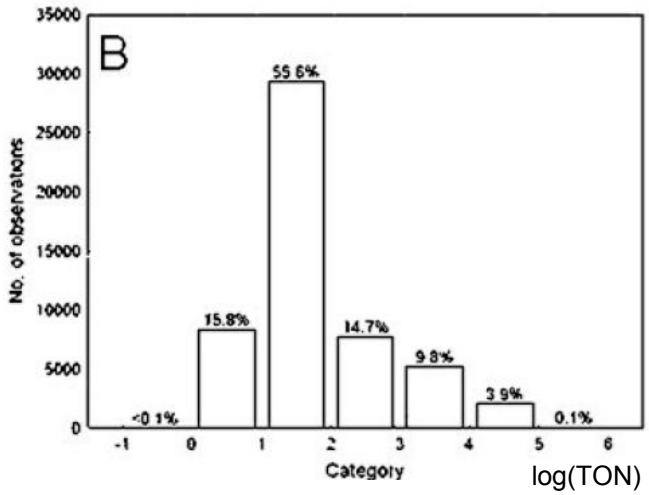
		Tree		LDA		ANN	
		true	false	true	false	true	false
TON	positive	92	12	92	13	89	6
	negative	307	1	306	1	313	4
TOF	positive	102	7	92	17	91	29
	negative	295	8	285	18	273	19

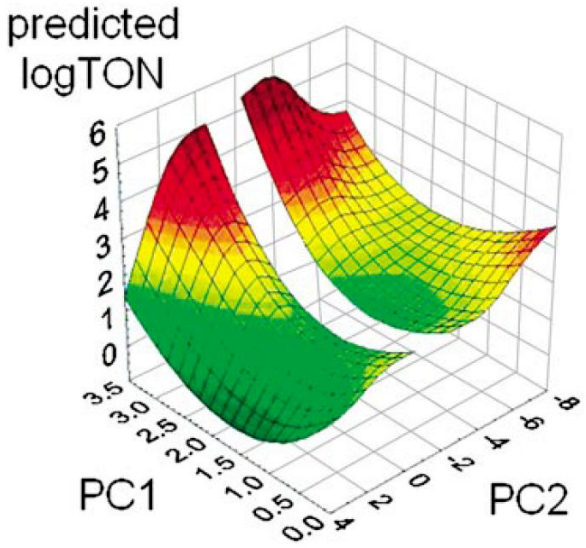
Models
- Tree : Classification Tree Models
- LDA : Linear-Discriminant-Analysis

Positive/Negative Threshold
- Log(TON) = 2 (TON = 100)
- Log(TOF) = 1 (TOF = 10)

Computational Screening of 60,000 Heck Reactions

61 PR^3 type ligands ∗ 4 olefins ∗ 4 aryl-X ∗ 5 catalyst precursors ∗ 4 solvents ∗ 3 Pd loadings = 58,560 conditions





PC1 is mainly correlated with Pd loadings and electronic descriptors of R^2.
PC2 is mainlys correlated with ligand electronic descriptors.

6. Summary

QSAR Approach

Advantages
- ·Short time calculation
- ·Easy to extract what is important
- ·Easy to search a good catalyst

Disadvantages
- ·Not based on reaction mechanism
- ·Only statistical estimation
- ·Many samples are necessary for good model.

Ab Initio Calculations of Transition States

Advantages
- ·Based on reaction mechanism.
- ·Many samples are not needed.

Disadvantages
- ·Long time calculation
- ·Difficult to predict what is important and good catalyst without intuition.