Catalyst Design Using Machine Learning

2020.10.10 Literature Seminar

B4 Junichi Taguchi

Contents

1. Introduction

2. Data-Driven Molecular Design in Asymmetric Catalysis

(*Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701**)**

3. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

(*Science* **2019,** *363, eaau5631***)**

Contents

1. Introduction

2. Data-Driven Molecular Design in Asymmetric Catalysis

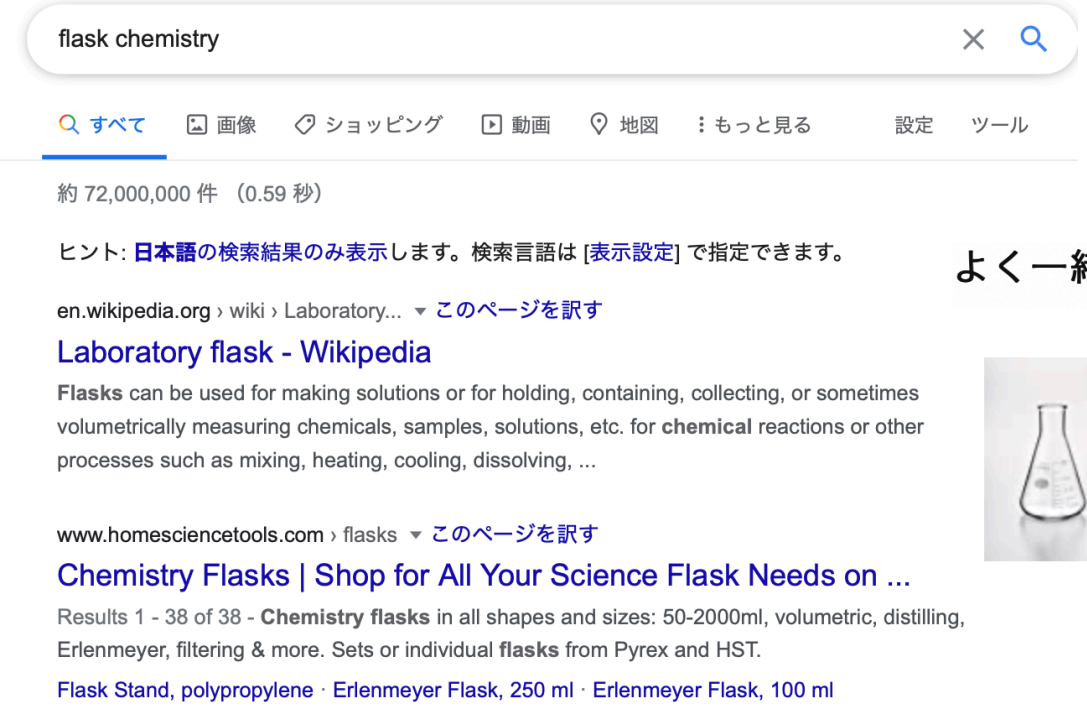
(*Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701)

3. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

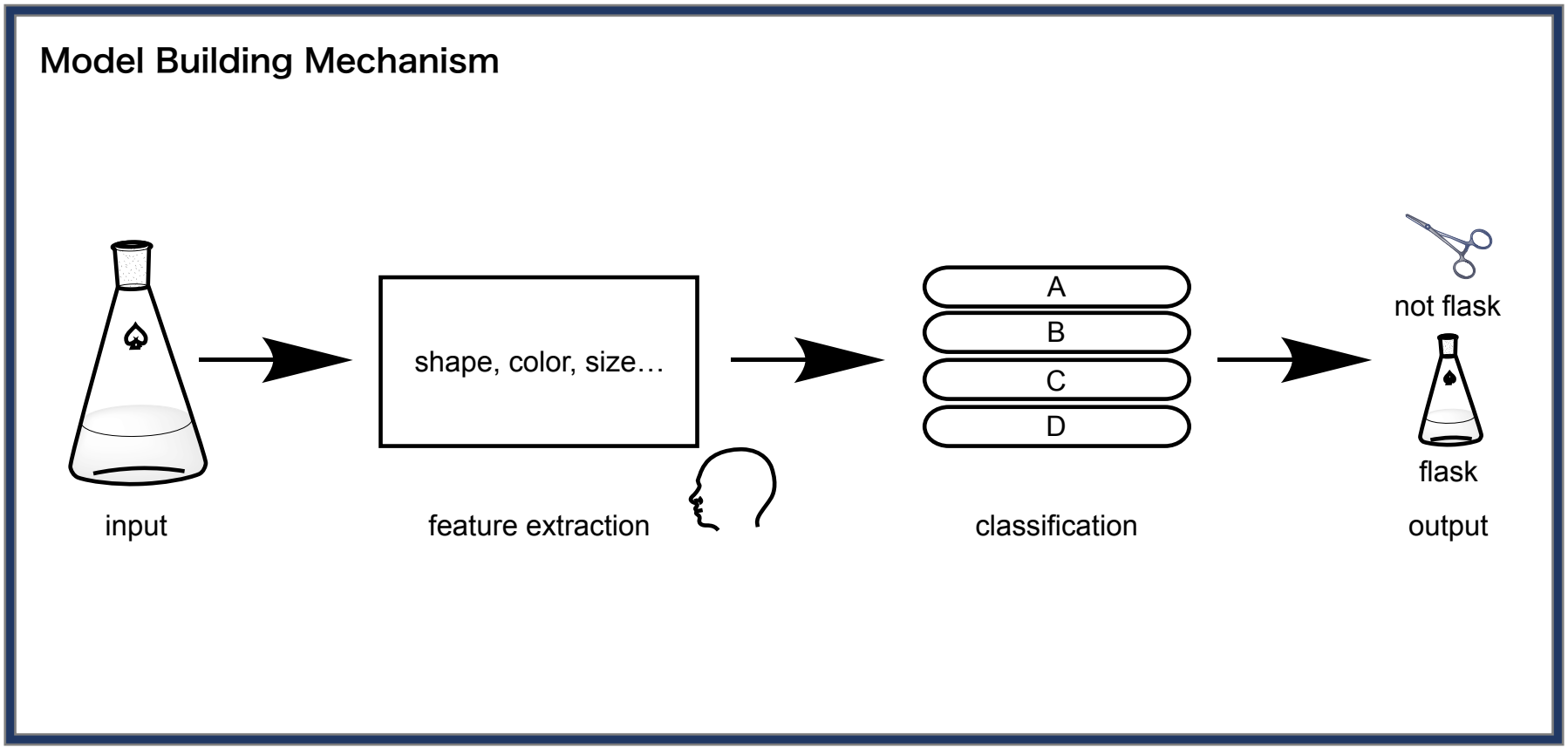
(*Science* **2019**, *363*, *eaau5631*)

What is Machine Learning (ML)

- Tool to find rules and patterns in the data
- Identification and prediction are the main uses
- It is likely to be more accurate than traditional methods



What is Machine Learning (ML)



See also 160916_LS_Yuki_NAKAGAWA and 180411_LS_Yusuke_Imamura

Catalyst Design Using Molecular Field Analysis (MFA)

MFA: A regression analysis between an objective variable and molecular fields calculated from 3D-molecular structures

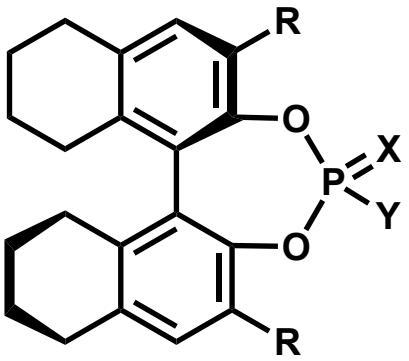
an objective variable = the enantiomeric ratios of products for the purpose of molecular design in asymmetric catalysis

Procedure

1. **Construction of an in silico library of catalysts**
2. **Calculation of molecular fields**
3. **Collection of the experimental data**
4. **Application of machine learning methods to generate models**

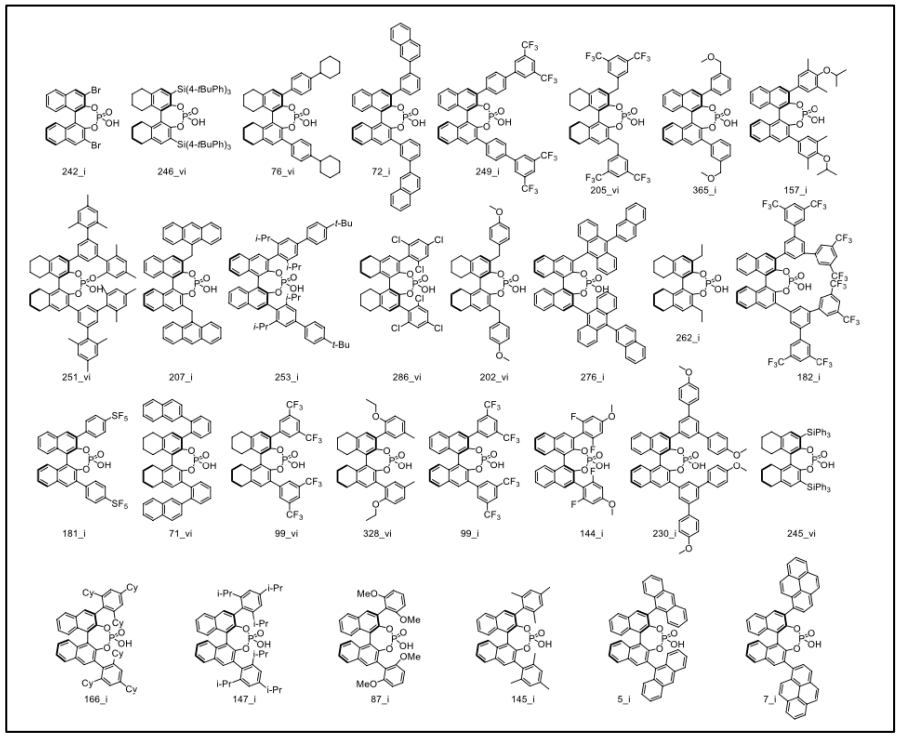
Catalyst Design Using MFA

step 1: Construction of an in silico library of catalysts



a particular scaffold

synthetic accessibility



in silico library

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Catalyst Design Using MFA

step 2: Calculation of molecular fields

regression analysis

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

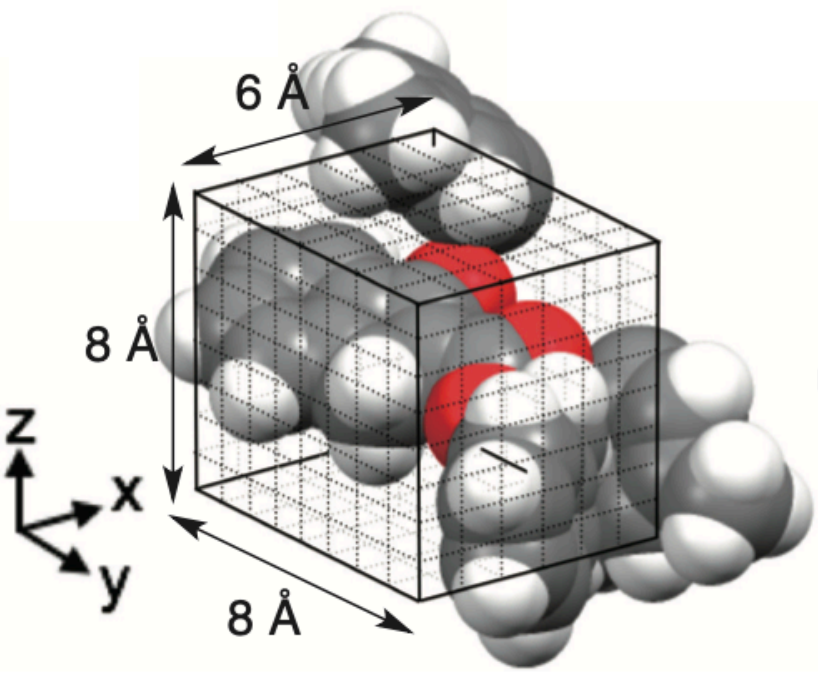
y：the logarithms of product enantiomeric ratios (k_{rel})
(i.e. $\Delta\Delta G^{\ddagger} = -RT\log k_{rel}$)

x_n： parameters for chemical properties of compounds
(i.e. “descriptor(記述子)”)

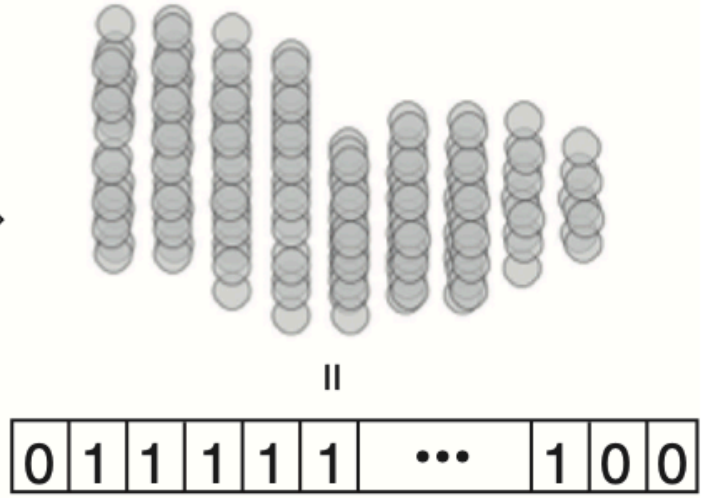
β： regression coefficient (Constants determined by performing a regression analysis)

Catalyst Design Using MFA

step 2: Calculation of molecular fields



placed in a grid space

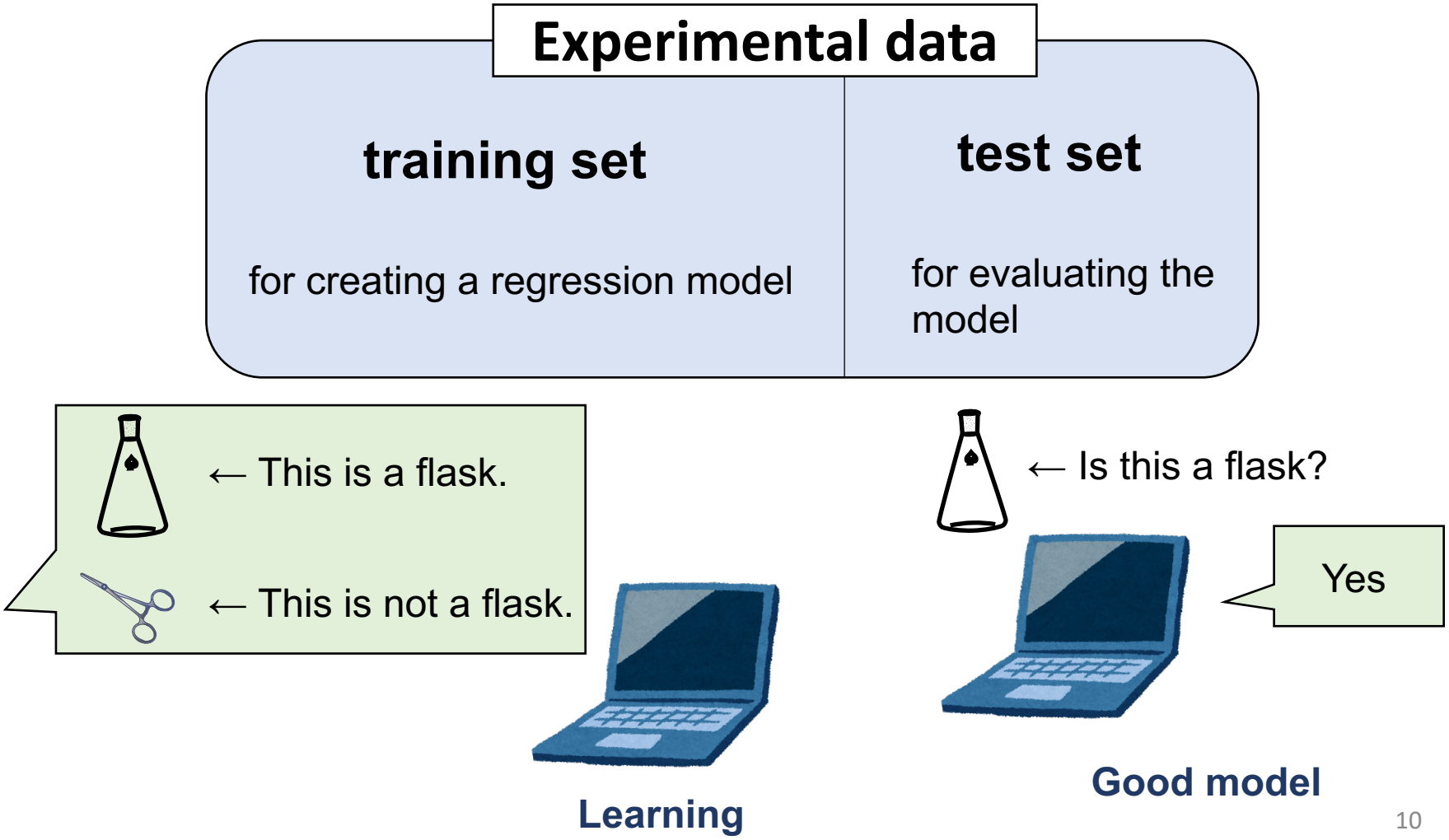


$\rightarrow x_n$

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Catalyst Design Using MFA

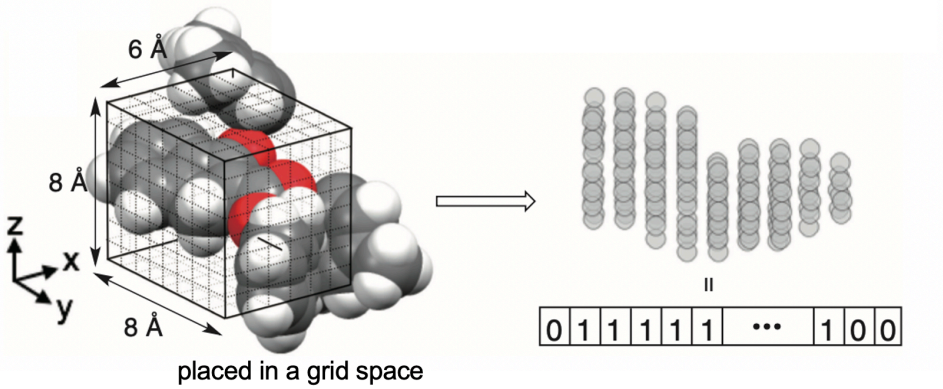
step 4: Application of machine learning methods to generate models



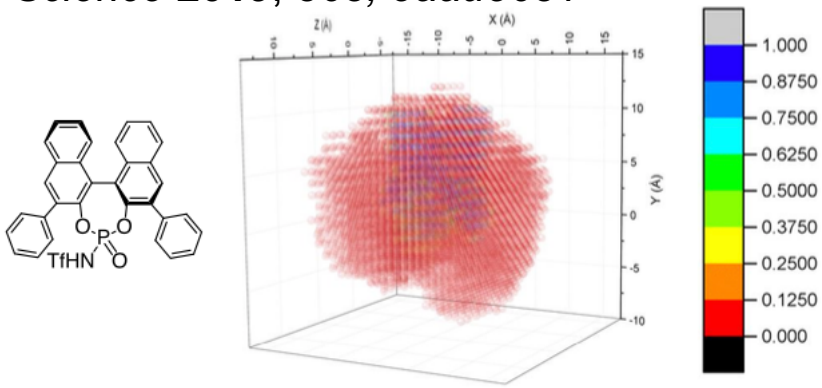
Key Points in This Lecture

- **Differences in molecular field analysis methods**

Bull. Chem. Soc. Jpn. **2019**, *92*, 1701



Science **2019**, *363, eaau5631*



- **Differences in the amount of the training data**

Bull. Chem. Soc. Jpn. **2019**, *92*, 1701

a small amount of reactions using theoretical approach

Science **2019**, *363, eaau5631*

a large amount of reactions using experimental approach

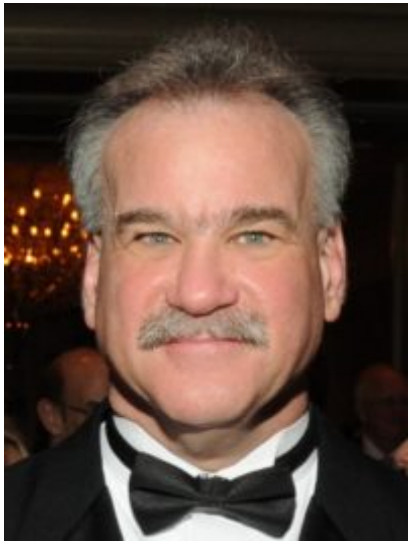
1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701
2) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Dr. Mikiko Sodeoka and Prof. Scott E. Denmark

1981	B.Pharm. @ Chiba University
1983	M.Pharm. @ Chiba University (Prof. T. Hino)
1989	Ph.D Pharmacy @ Chiba University (Prof. M. Nakagawa)
1992-	Research Associate @ University of Tokyo (Prof. M. Shibasaki)
1999-	Associate Professor @ University of Tokyo
2000-	Professor @ Tohoku University
2006-	Chief Scientist, Synthetic Organic Chemistry Laboratory, RIKEN

Research Interests :

transition metal chemistry/organic synthetic chemistry/chemical biology



1975	S.B. @ Massachusetts Institute of Technology
1980	D.Sc.Tech @ ETH Zurich (Prof. Albert Eschenmoser)
1980-	Assistant Professor @ University of Illinois
1986-	Associate Professor @ University of Illinois
1987-	Professor @ University of Illinois

Research Interests :

synthetic organic chemistry/invention of new organic reactions/
total synthesis of natural products

1) https://chemistry.illinois.edu/sdenmark
2) https://www.jst.go.jp/erato/sodeoka/english/profile/index.html

Contents

1. Introduction

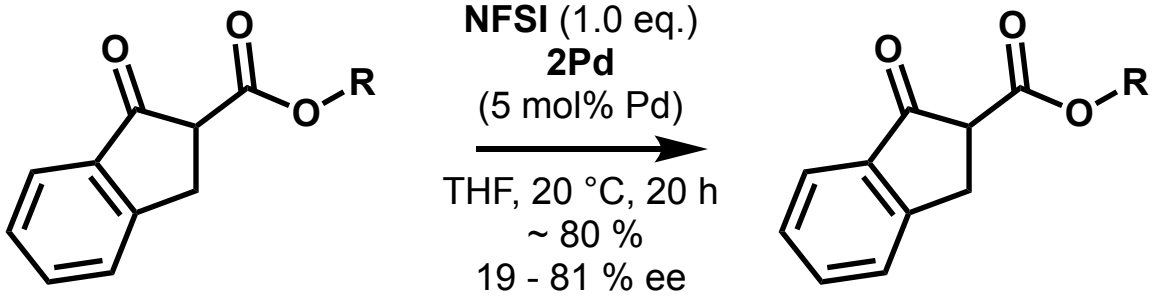
2. Data-Driven Molecular Design in Asymmetric Catalysis

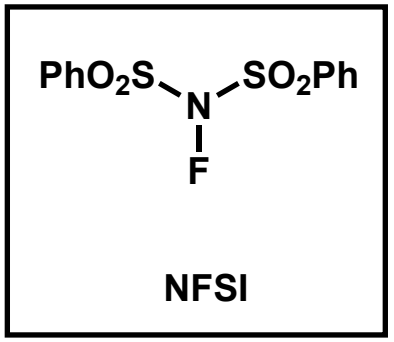
(*Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701**)**

3. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

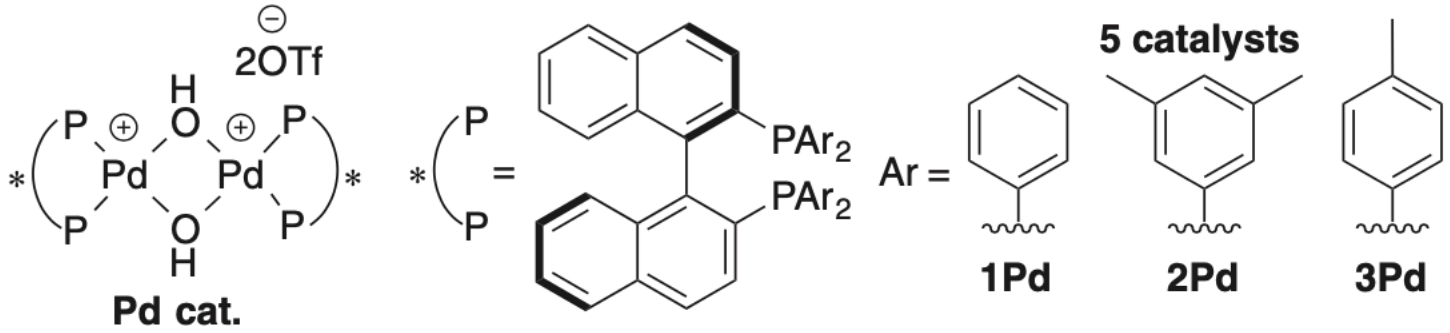
(*Science* **363**, *eaau5631*)

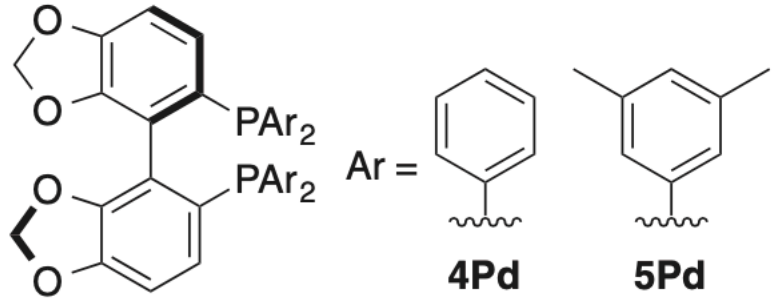
Model Reaction System (step 1)





R = Me, Et, Bn, *i*Pr, Cy, *t*Bu (6 substrates)





$$6\ substrates \times 5\ catalysts = 30\ reaction$$

24 reactions were used as training data

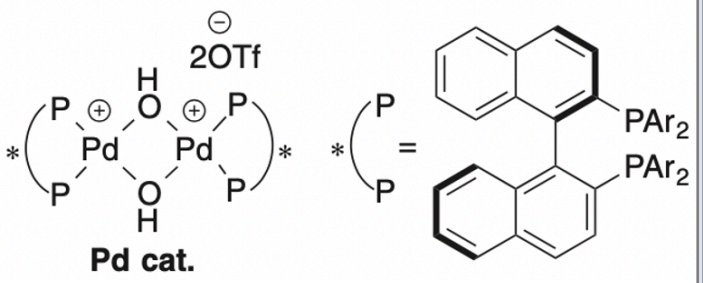
1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn* **2019**, *92*, 1701
2) Hamashima, Y.; Yagi, K.; Takano, H.;Tamás, L.; Sodeoka, M. *J. Am. Chem. Soc.* **2002**, *124*, 14530

Calculation of Molecular Fields (step 2)

Most MFA reported to date

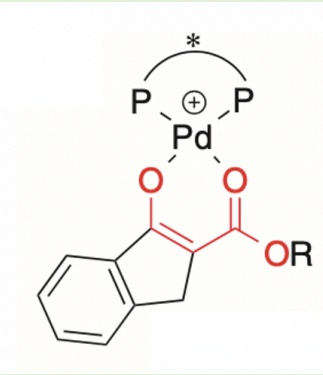
Molecular structures without substrates in calculations of molecular fields were employed for MFA

→There are no example of molecular design to improve enantioselectivity



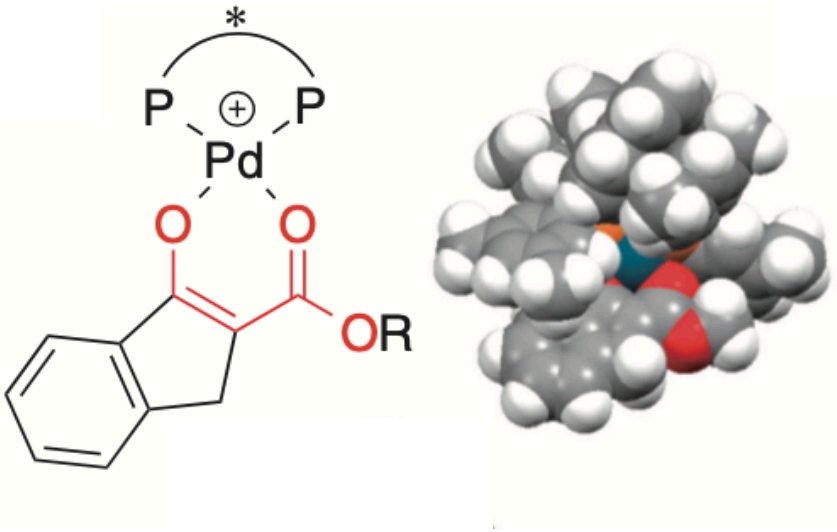
In this study

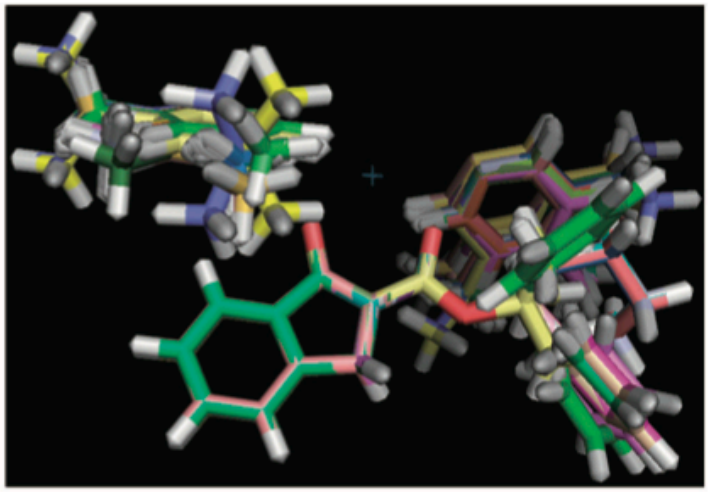
Intermediate structures in an asymmetric induction step were employed for MFA



1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Calculation of Molecular Fields (step 2)





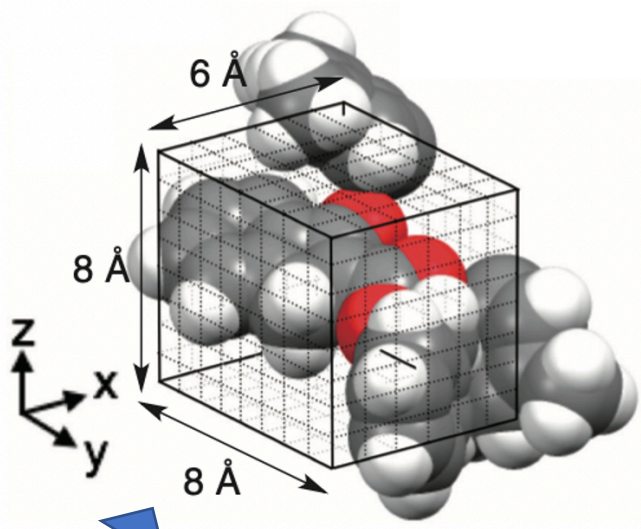
Optimization of Pd-enolate structures (B3LYP/LANL2DZ(Pd) and 6-31G(d) level)

Alignment of the set of intermediates

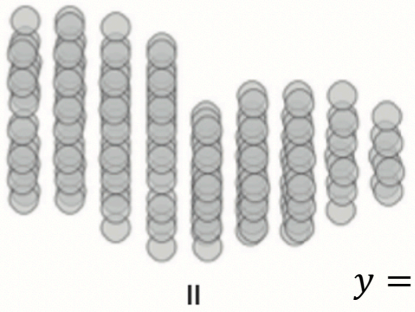
Atoms except for the β-ketoester and equatorial Ar-groups on the ligands were removed.

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Calculation of Molecular Fields (step 2)



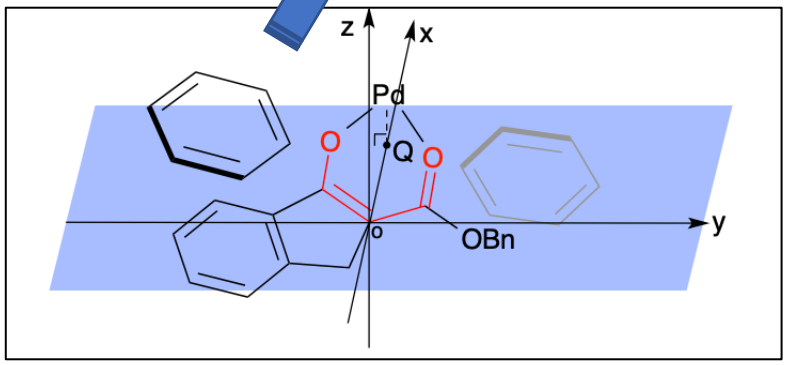
placed in a grid space



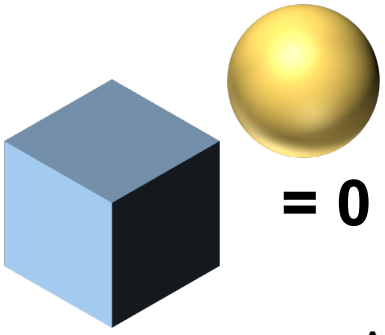
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

0	1	1	1	1	1	•••	1	0	0

$\rightarrow x_n$

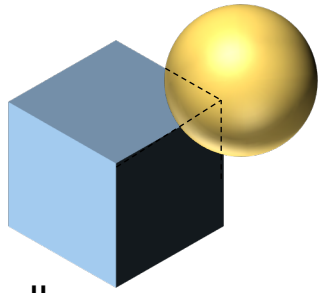


indicator field



= 0

A van der Waals radii of an atom

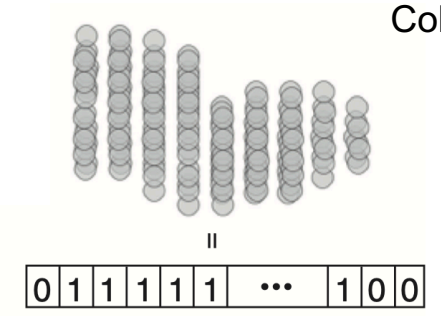


= 1

A unit cell

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Application of ML Methods to Generate Models (step 4)



Columns in the descriptor matrix that exhibited small deviations were removed

$x_1, x_2, \cdots, x_{61}$

61 dimensional descriptors

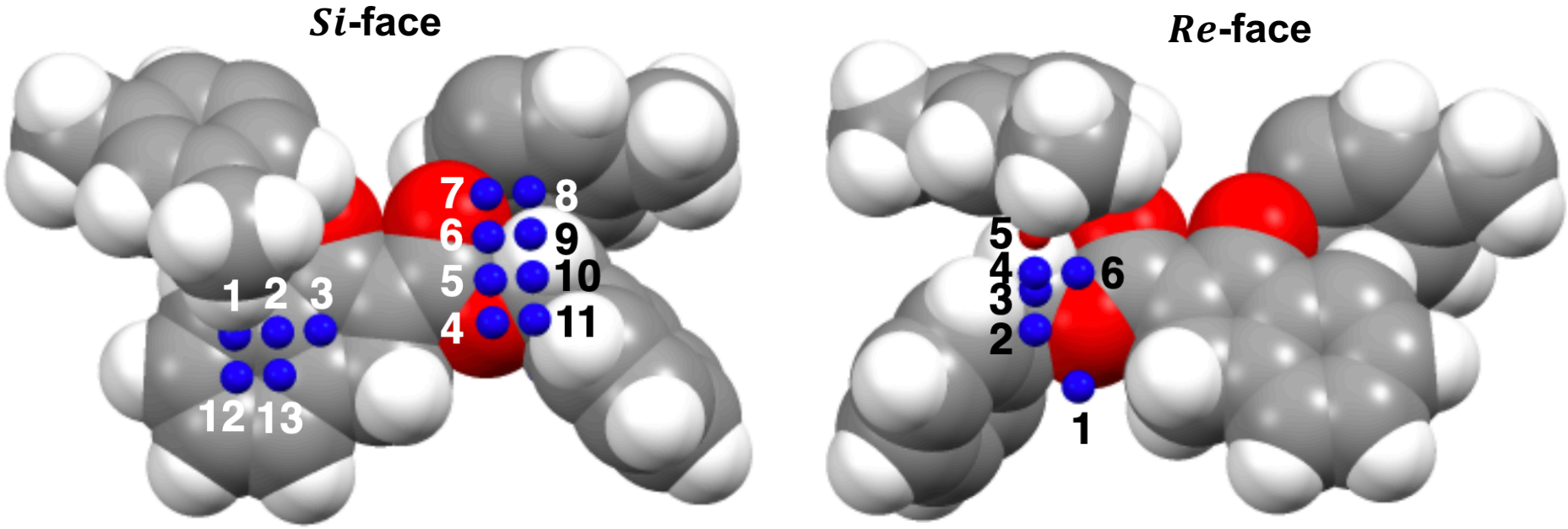
Elastic Net regression

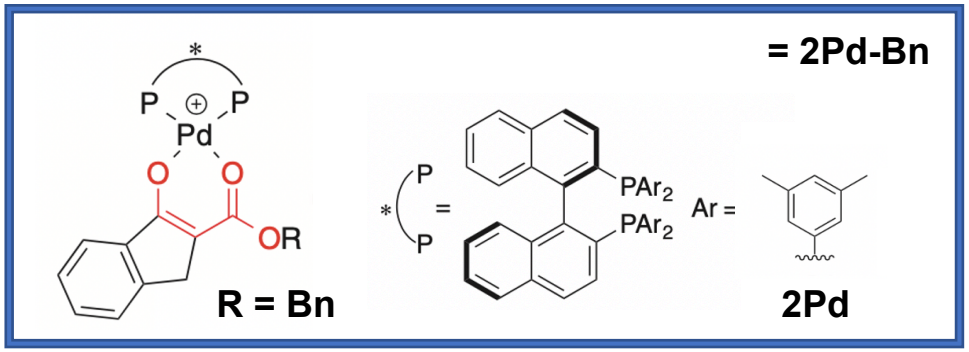
19 important descriptors were extracted

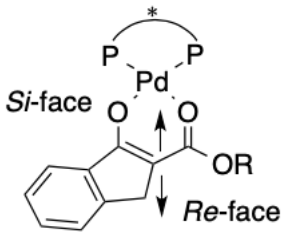
= 19 important unit cells  **were extracted**

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Application of ML Methods to Generate Models (step 4)





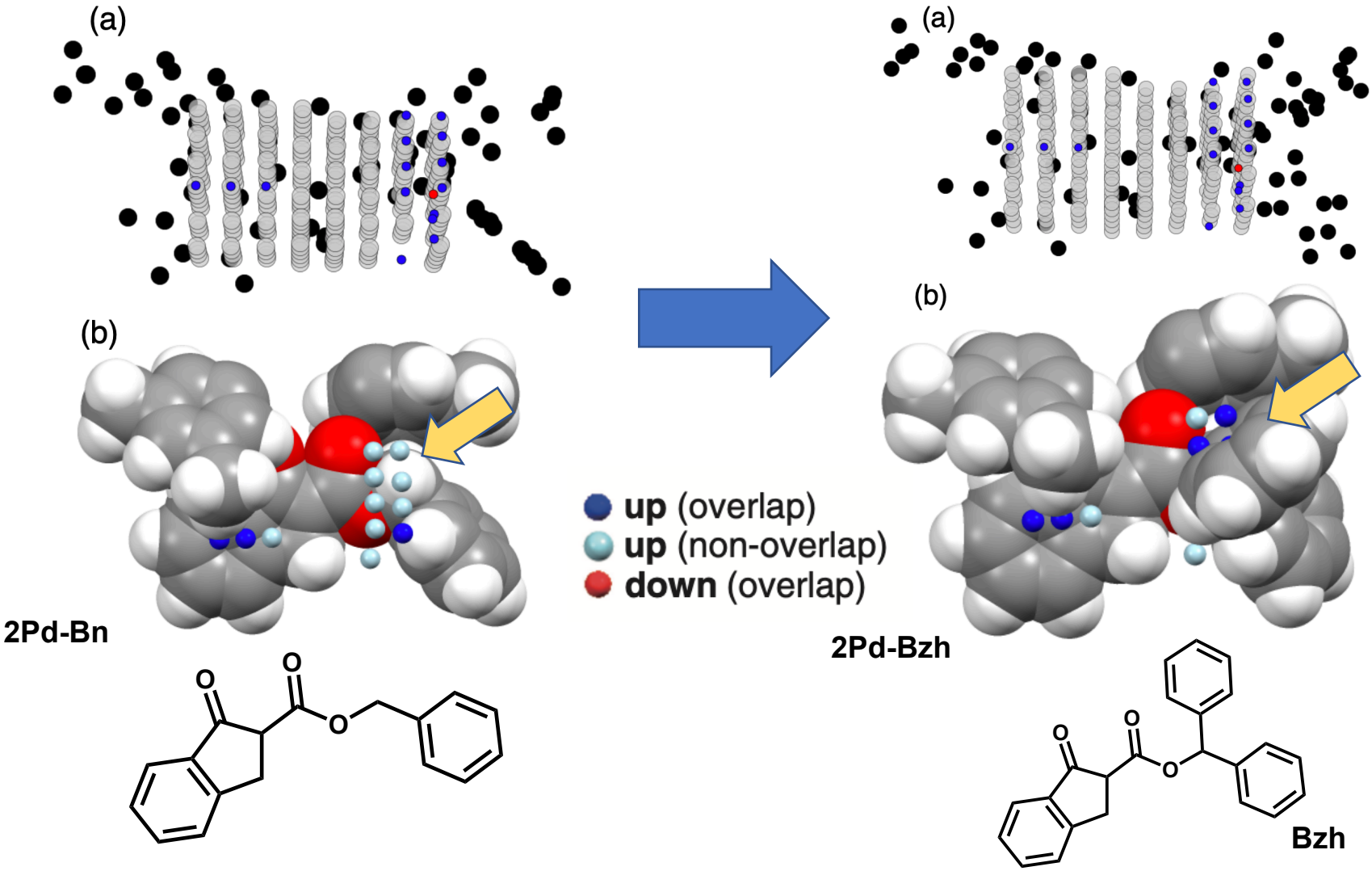


up (overlap)

down (overlap)

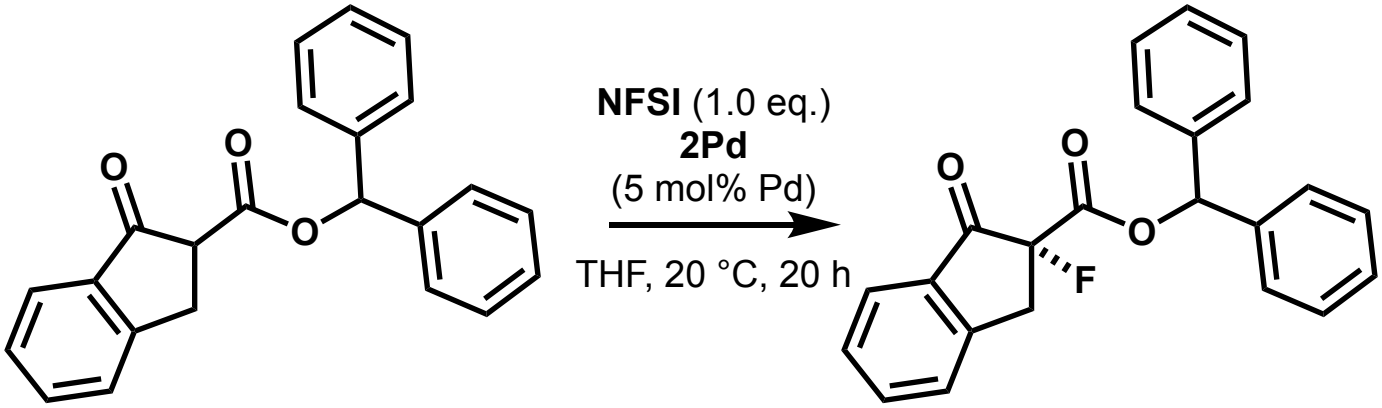
1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

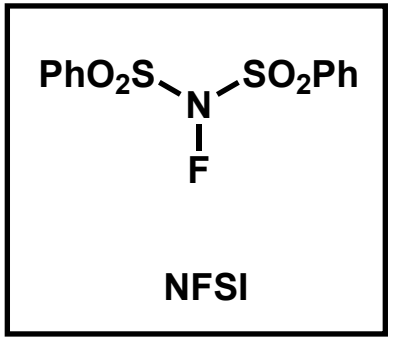
Molecular Design to Improve Enantioselectivity



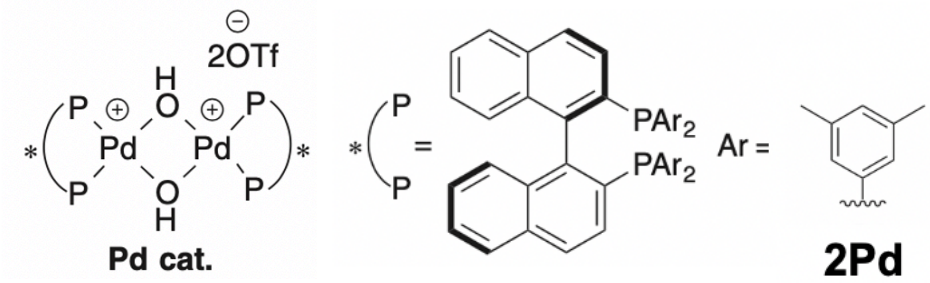
1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Molecular Design to Improve Enantioselectivity





calculated ΔΔG‡ = 3.6 kcal/mol (about 99 % ee)
Experimental result : 94 % ee

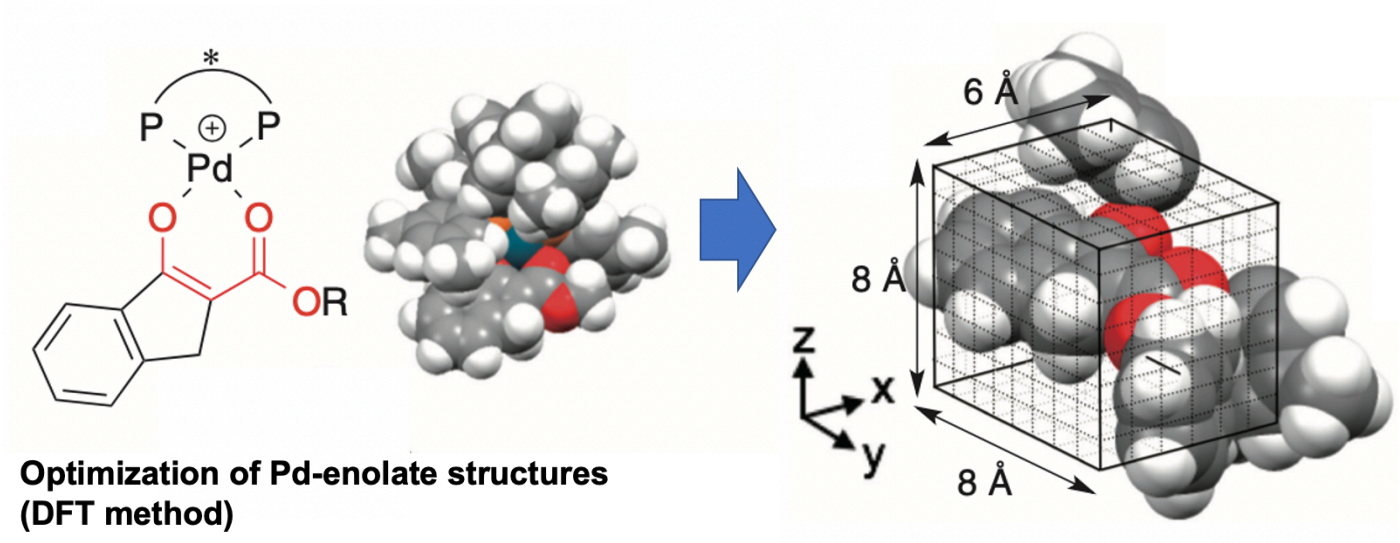


Better than any of the data in the training set

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Short Summary

- **Molecular field analysis methods:**
  Intermediate structures in an asymmetric induction step were employed for MFA



Optimization of Pd-enolate structures (DFT method)

- **The amount of the training data:**
  A little amount of reactions (24 reactions) were used as a training data

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Contents

1. Introduction

2. Data-Driven Molecular Design in Asymmetric Catalysis

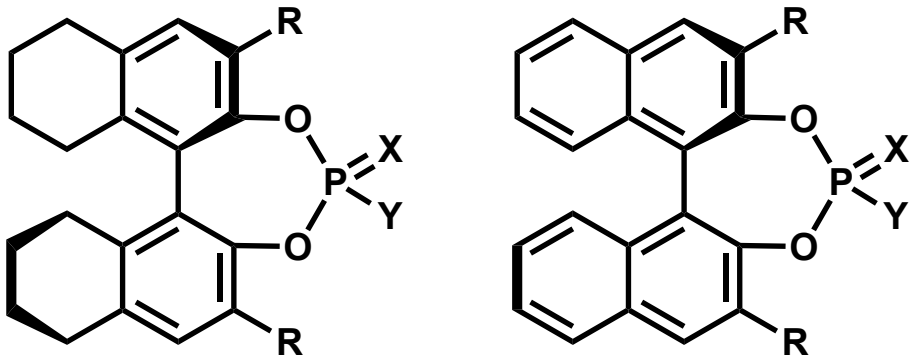
(*Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701)

3. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

(*Science* **2019,** *363, eaau5631***)**

The catalyst scaffold (step 1)

target: The BINOL (1,1'-bi-2-naphthol)-derived family of chiral phosphoric acids



- **synthetic accessibility**
- **ease of diversification by installation of an array of substituents**
- **the acidity of the phosphoryl group can be unsaturated**
- **the backbone can be unsaturated or saturated**
- **can be used for a vast number of synthetically useful reactions**

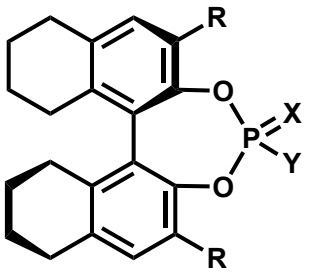
806 chiral phosphoric acid catalysts in the silico library

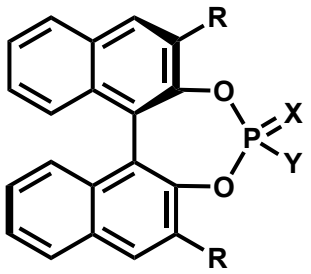
1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Calculation of Molecular Fields (step 2)

Most MFA reported to date

Molecular structures without substrates in calculations of molecular fields were employed for MFA

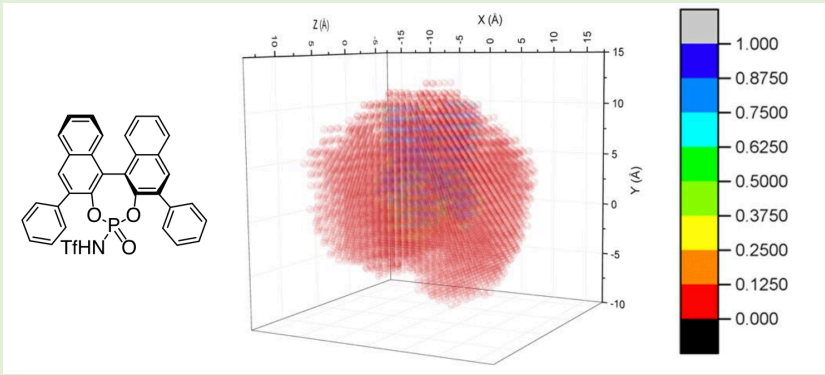




→There are no example of molecular design to improve enantioselectivity

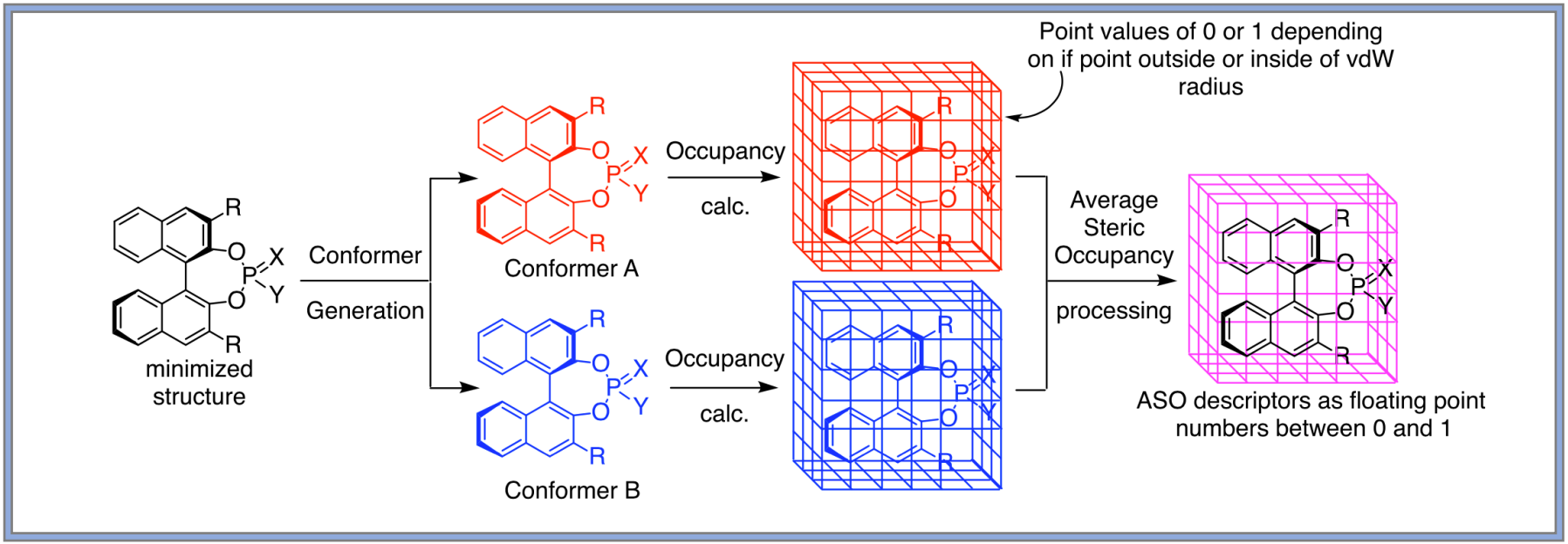
In this study

Average Steric Occupancy (ASO) were employed for MFA

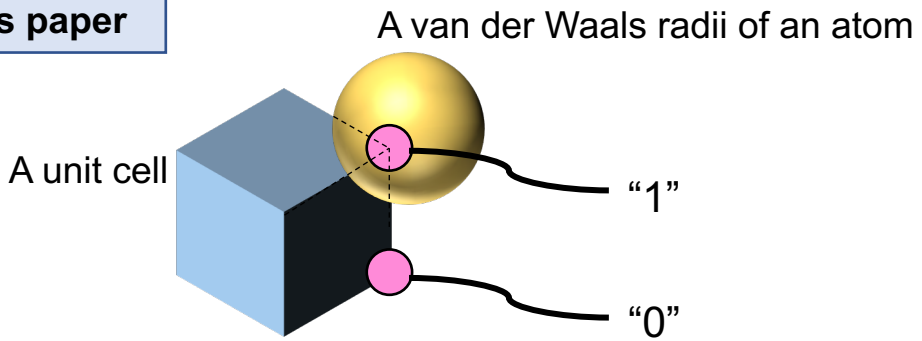


1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701
2) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Average Steric Occupancy (ASO) (step 2)



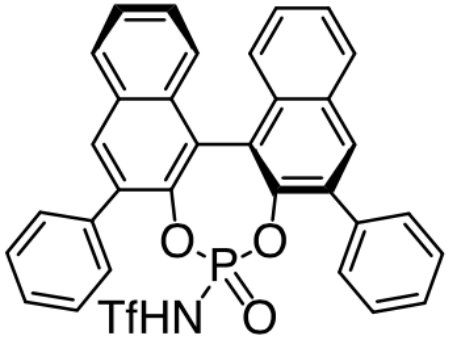
The indicator field in this paper

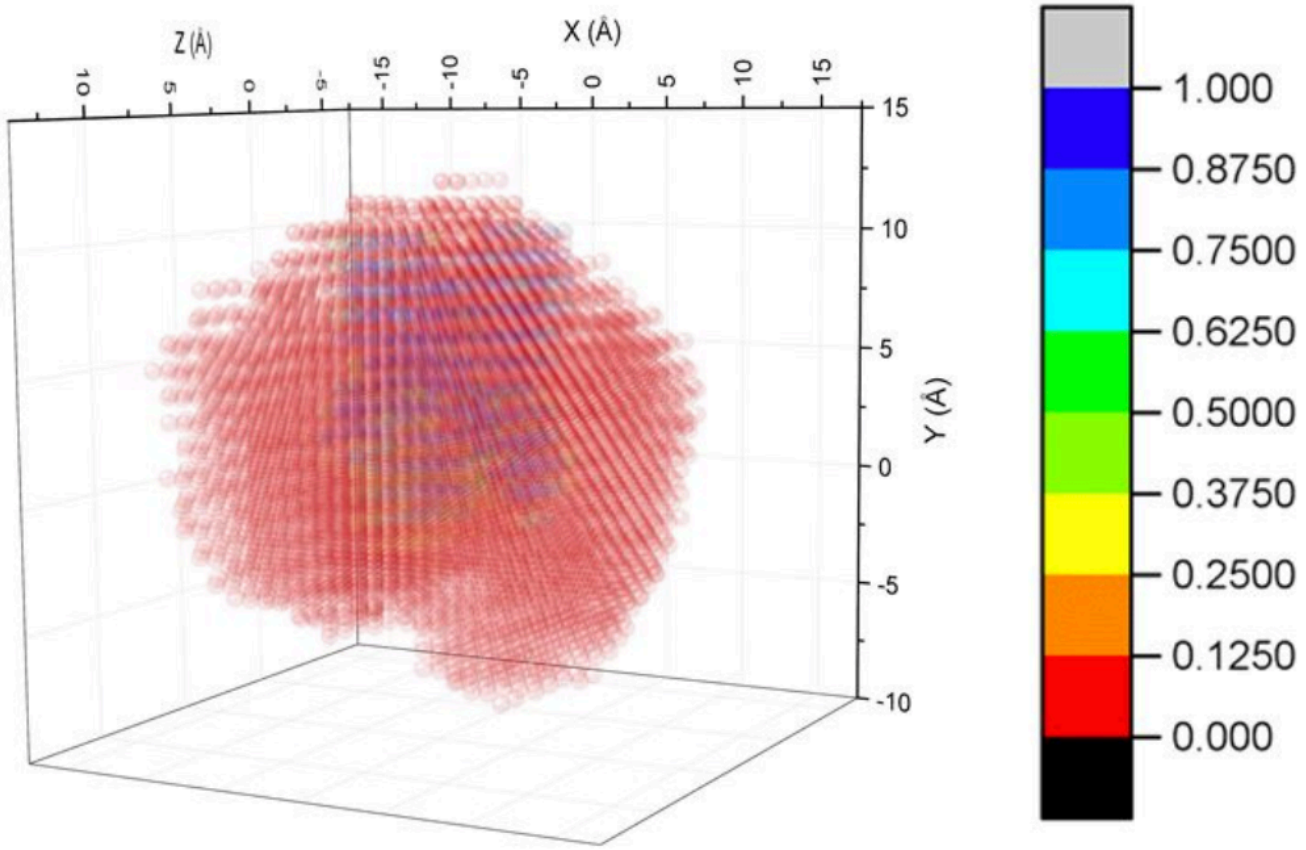


The grid points that included the van der Waals radii of any atoms were counted as 1, or were otherwise counted as 0.

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Average Steric Occupancy (ASO) (step 2)





ASO and electronic descriptors for reactants and products (+catalyst)
→individual reaction profiles including substrate properties.

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Universal Training Set (UTS) (step 3)

Conventional methods

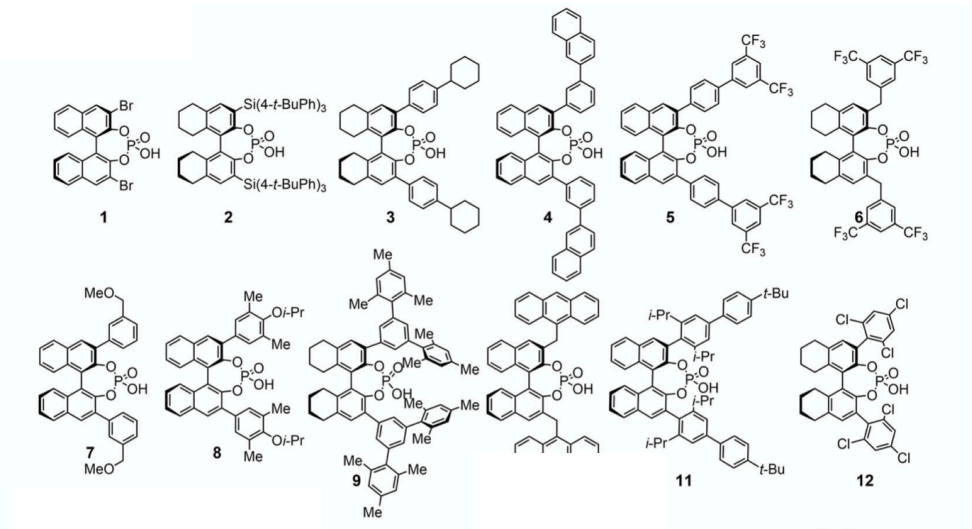
- A training set was chosen randomly from experimental data.
- There is a risk that the model would not be sufficiently accurate, depending on the variability of the results used for the training set.

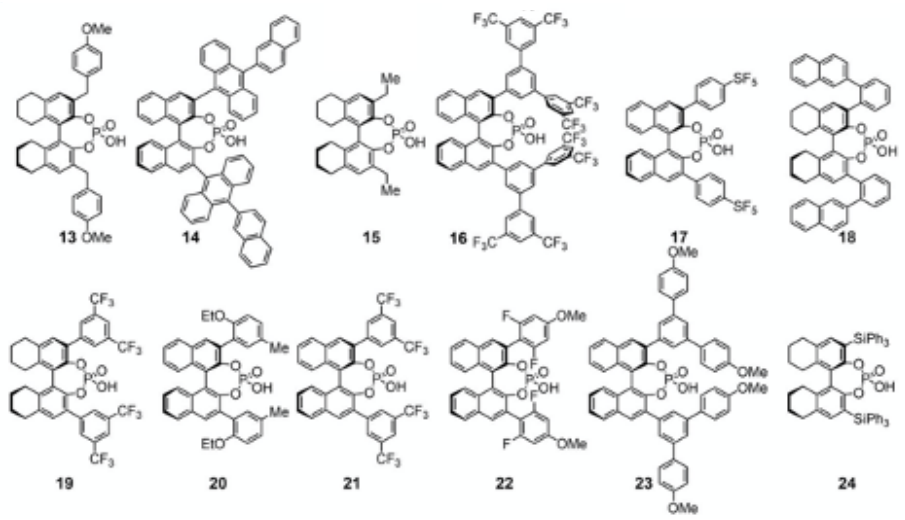
UTS

- Catalysts from uniform regions of feature space are sampled
- It can be used to optimize any reaction that can be catalyzed by that catalyst type.

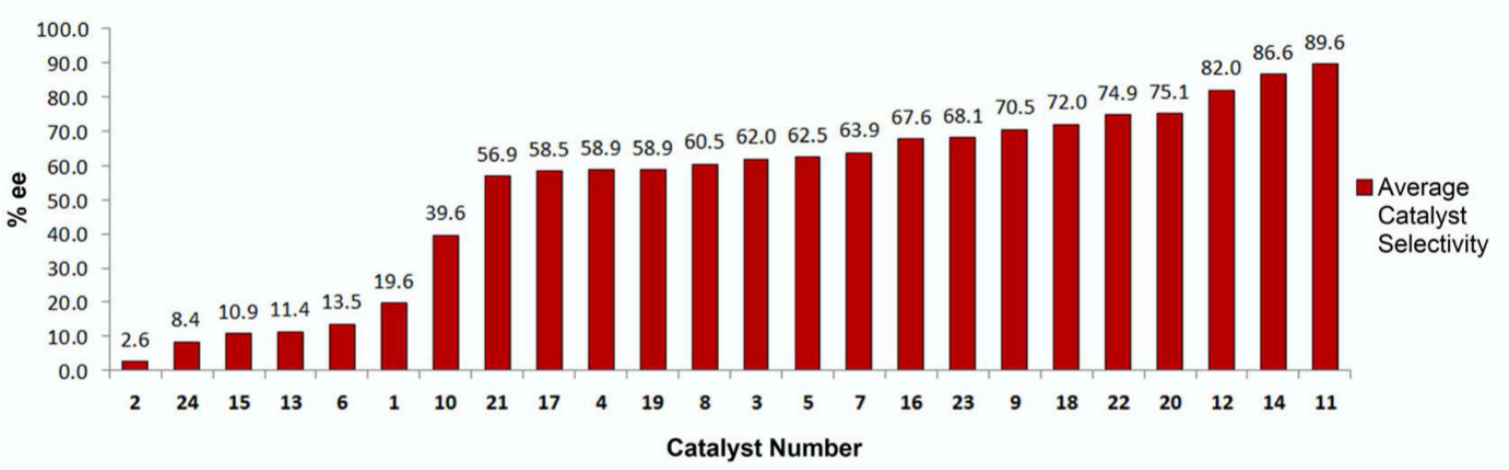
1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Universal Training Set (UTS) (step 3)



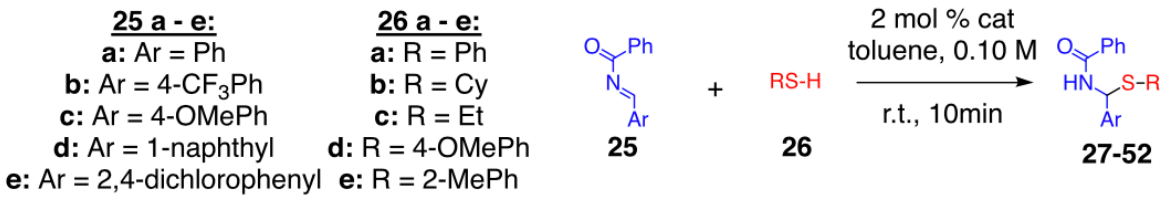


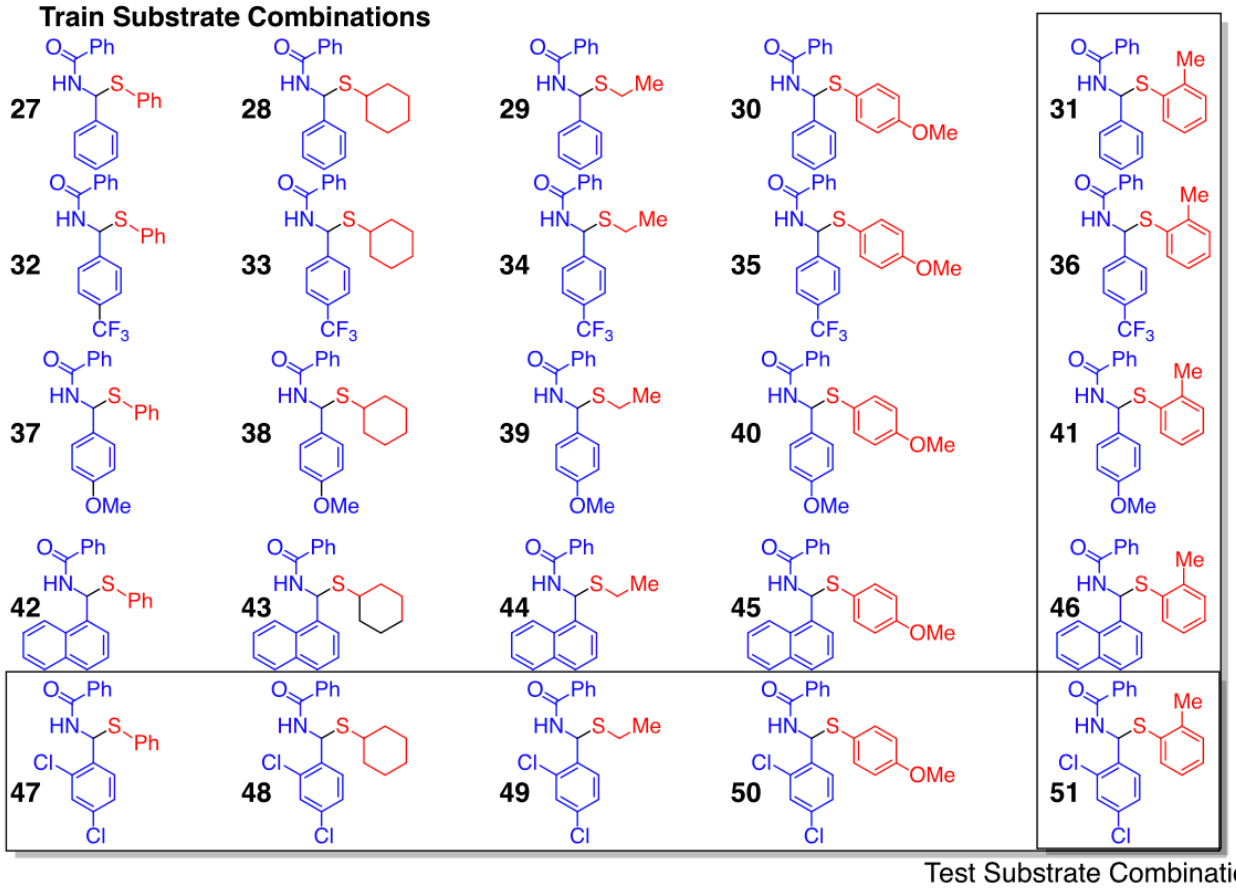
24 catalysts



1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Model Reaction (step 3)



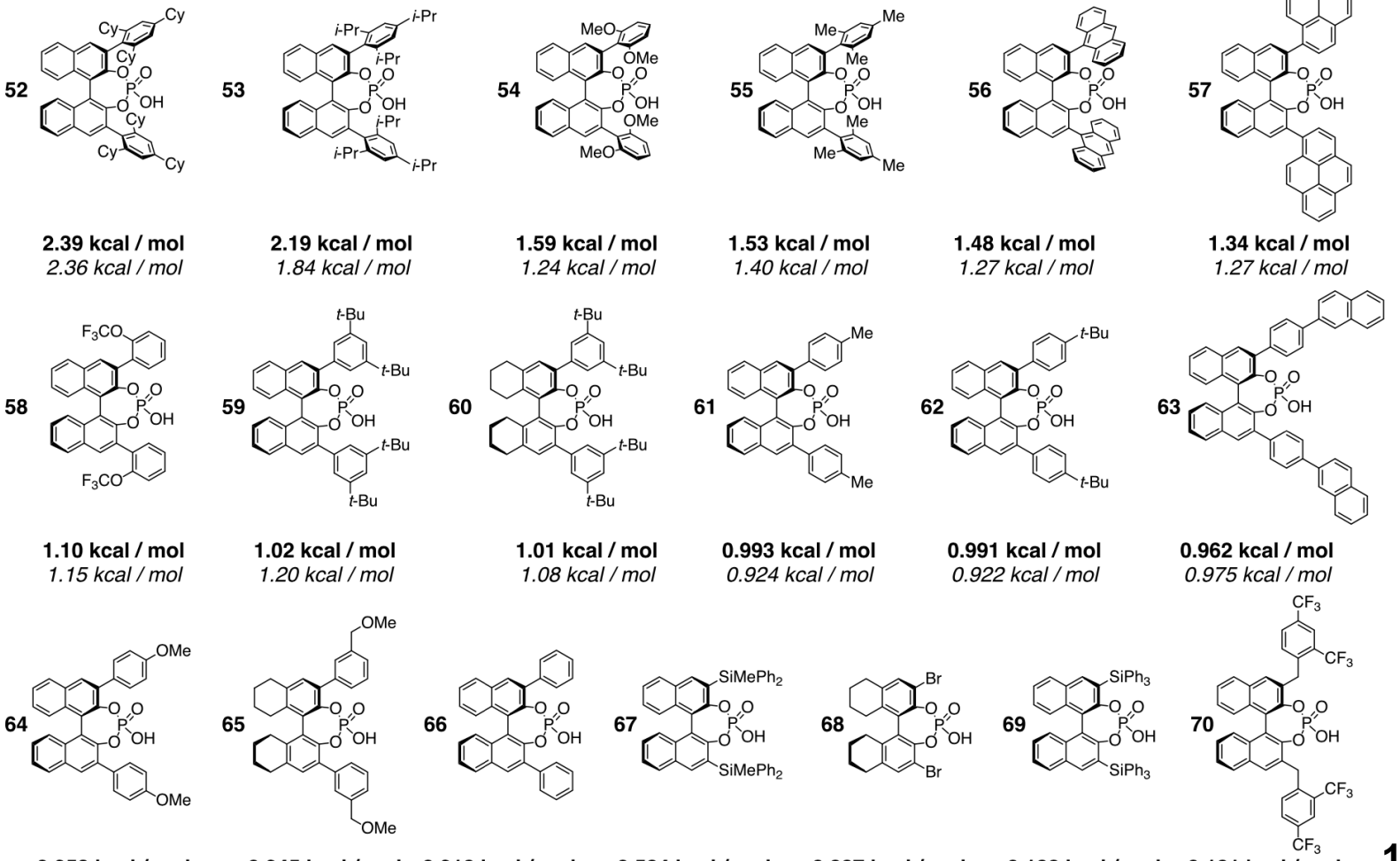


25 combinations

1) Henle, J. J.; Zahrt, A. E.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. *J. Am. Chem. Soc.* **2020**, *142,* 11578
2) Ingle, G. K.; Mormino, M. G.; Wojtas, L.; Antilla, J. C. *Org. Lett.* **2011**, *13*, 4822

Test Catalysts with Averages for All Substrate Combinations (step 3)

Observed
Predicted



19 catalysts

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Modeling Study

Could this tool be used to predict the results of either new substrate combinations or new catalysts?

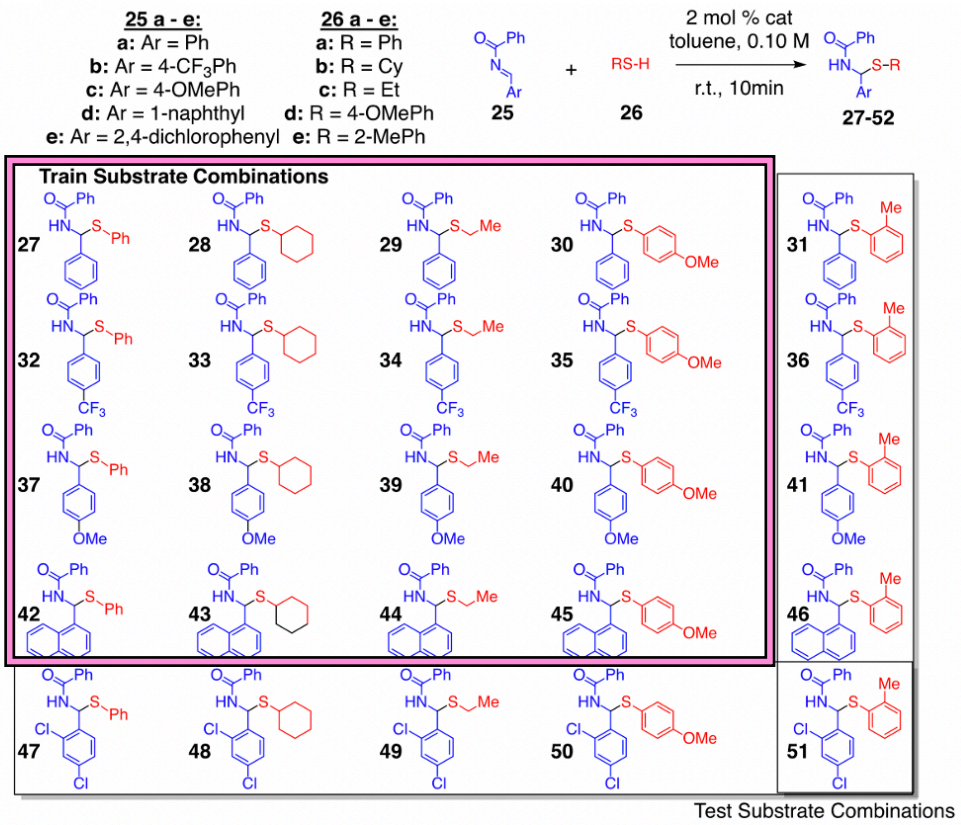
training set

16 reactions

×

24 catalysts (UTS)

=384 data



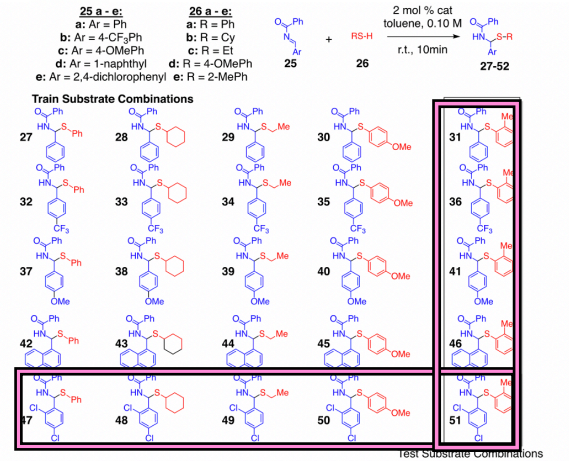
1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Modeling Study

Could this tool be used to predict the results of either new substrate combinations or new catalysts?

test set

a) new substrate



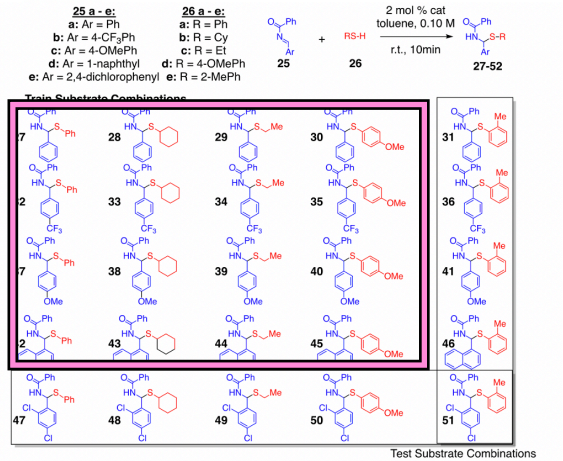
9 combinations

×

UTS (24 catalysts)

= 216 reactions

b) new catalyst



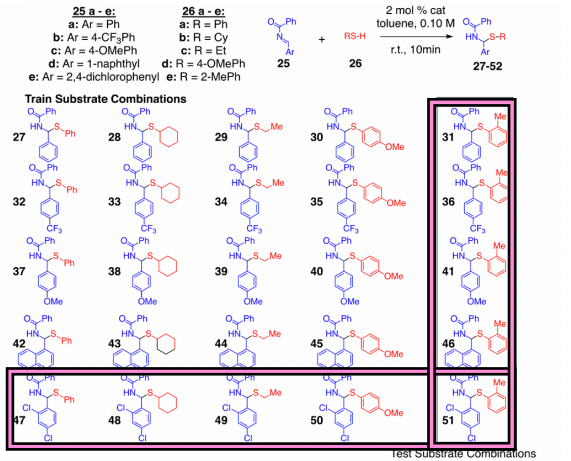
16 combinations

×

Test catalysts (19 catalysts)

= 304 reactions

c) new substrate & new catalyst



9 combinations

×

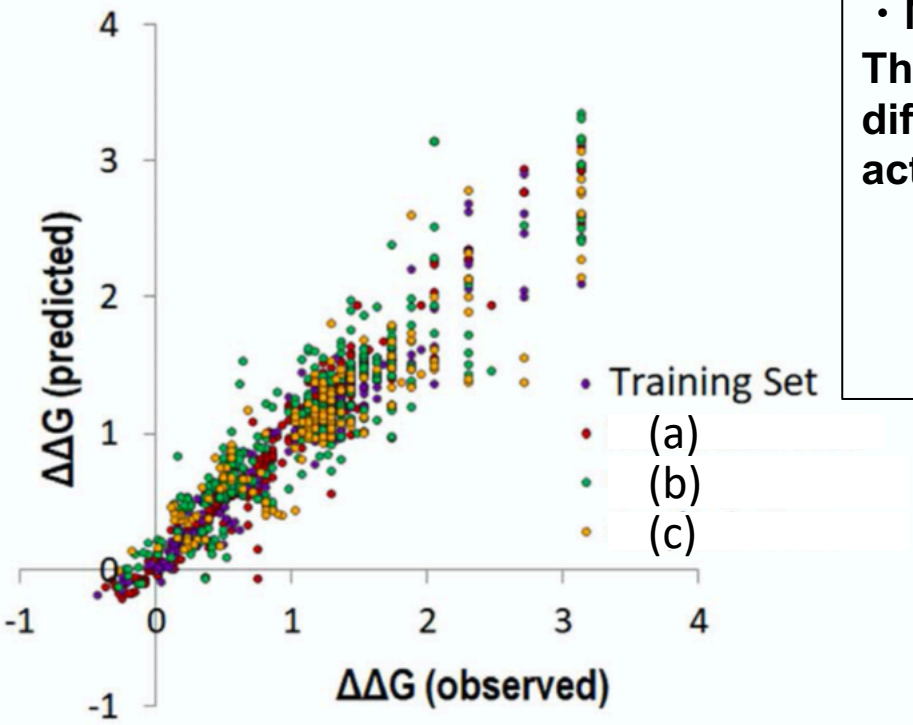
Test catalysts (19 catalysts)

= 171 reactions

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Modeling Study

Could this tool be used to predict the results of either new substrate combinations or new catalysts?



the support vector machines method was used (second-order polynomial kernel, q^2 = 0.748 by k-fold cross validation)

・Mean Absolute Deviation (MAD)：
The average of the absolute value of the difference between the predicted and actual values

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

a) MAD: 0.161 kcal/mol
b) MAD: 0.211 kcal/mol
c) MAD: 0.236 kcal/mol

→ The model made good predictions

$\Delta\Delta G^{\ddagger}$ = 3.0 kcal/mol → about 99 % ee
0.5 kcal/mol → about 40 % ee

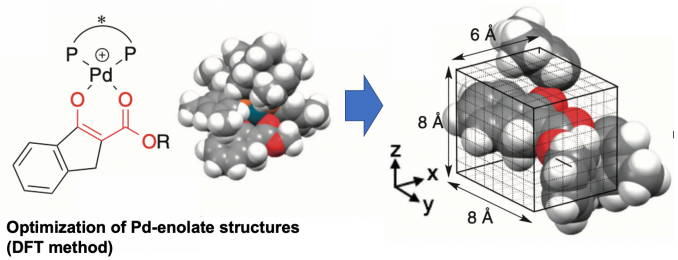
1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Summary

The main differences between two methods

Bull. Chem. Soc. Jpn. **2019**, *92*, 1701

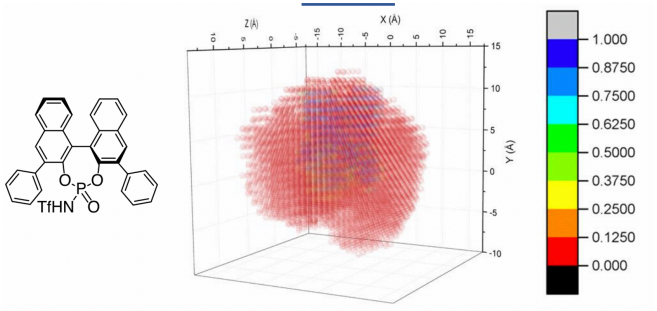
Intermediate structures in an asymmetric induction step



Optimization of Pd-enolate structures (DFT method)

Science **2019**, *363, eaau5631*

ASO



calculation of molecular fields

generating models

A little amount of reactions (24 reactions)

training data

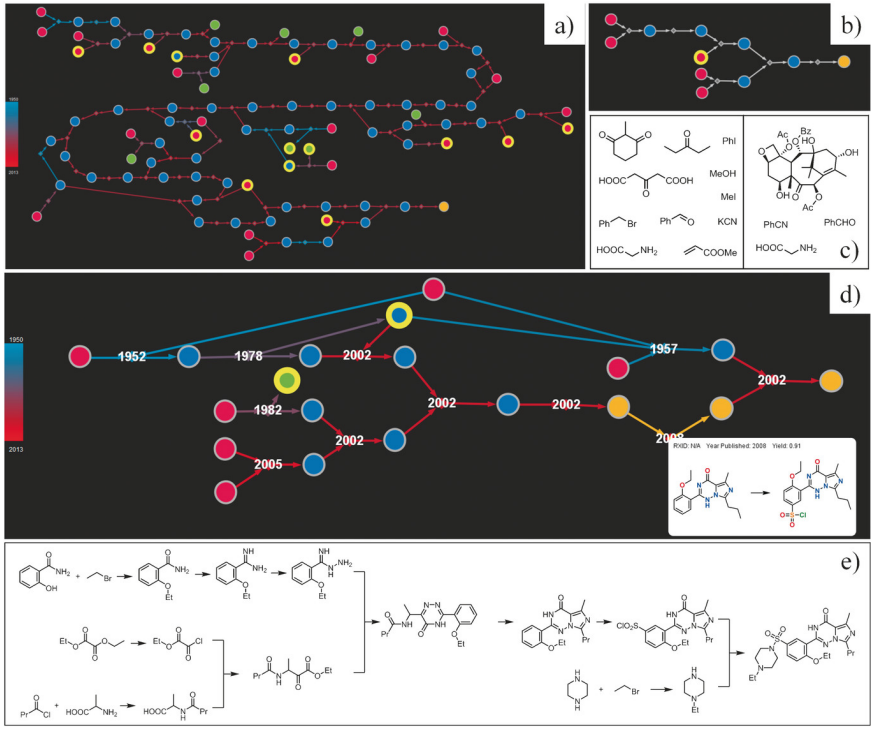
A large amount of reactions (using "UTS")

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701
2) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

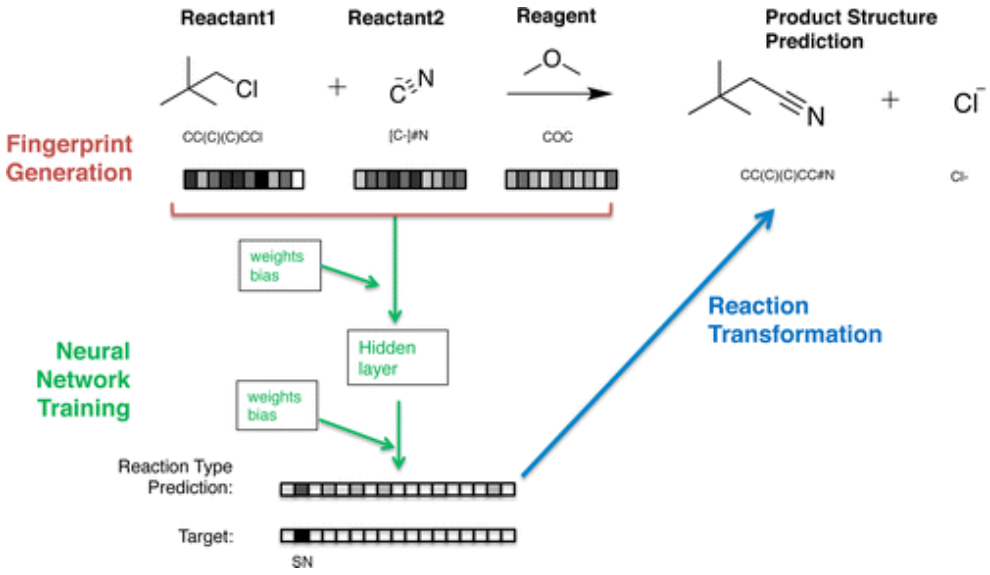
Appendix

ML and Organic Chemistry

Computer-assisted synthetic planning



Prediction of organic reaction outcome



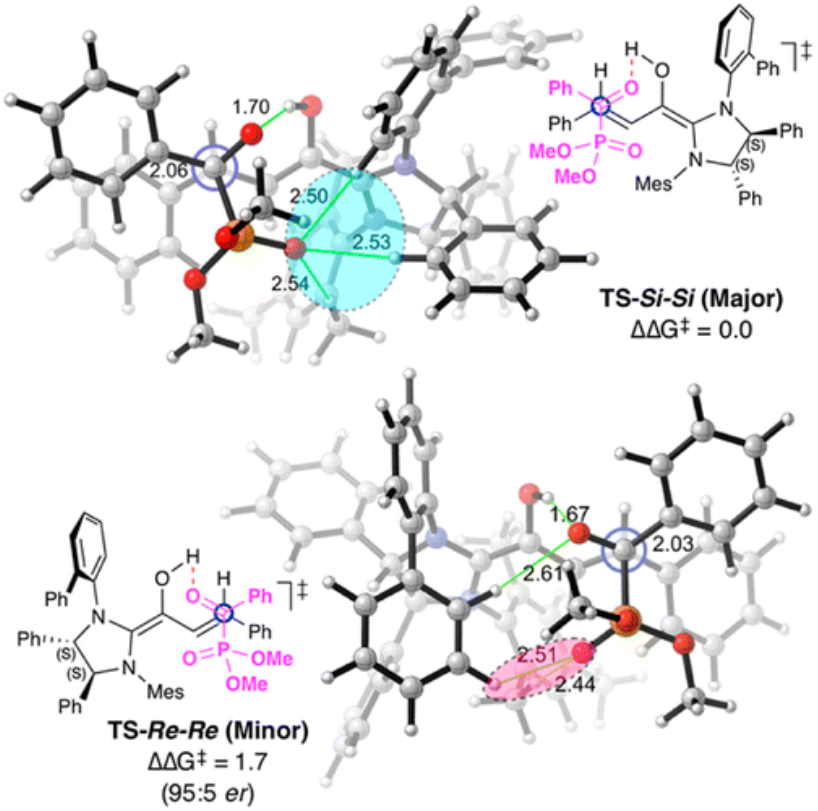
See also 160916_LS_Yuki_NAKAGAWA and 180411_LS_Yusuke_Imamura

1) Szymkuć, S.; Gajewska, E.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. *Angew. Chem. Int. Ed.* **2016**, *55*, 5904.

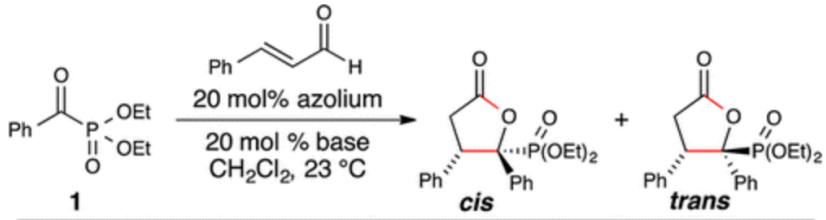
2) Wei, J. N.; Duvenaud, D.; Aspuru-Guzic, A. *ACS Cent. Sci.* **2016**, *2*, 725.

Conventional Process of Catalyst Design

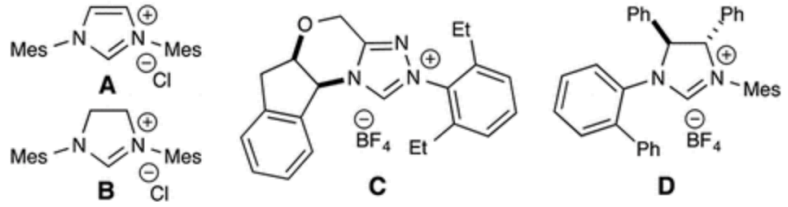
DFT calculation



Catalyst/Substrate Optimization

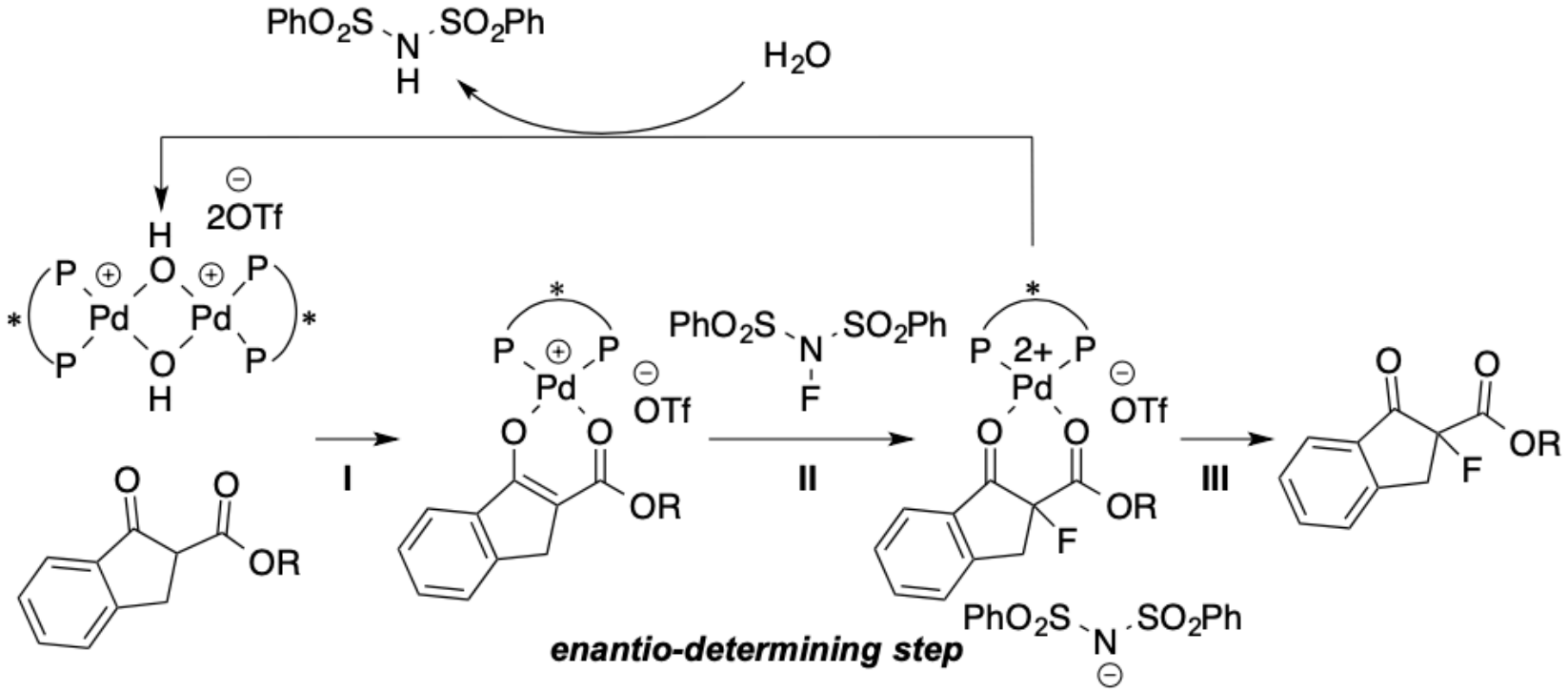


entry	azolium	base	% yield[a]	*cis*:*trans*[b]	er[c]
1	A	DBU	35	1:13	-
2	B	DBU	70	1:4.6	-
3	C	DBU	20	1.2:1	67:33
4	D	DBU	94	5.8:1	85:15
5	D	MTBD	79	1.8:1	90:10



1) Jang, K. J.; Hutson, G. E.; Johnson, R. C.; McCusker, E. O.; Cheong, P. H. -Y.; Scheidt, K.A. *J. Am. Chem. Soc.* **2014**, *136*, 76.

Proposed Reaction Mechanism



1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

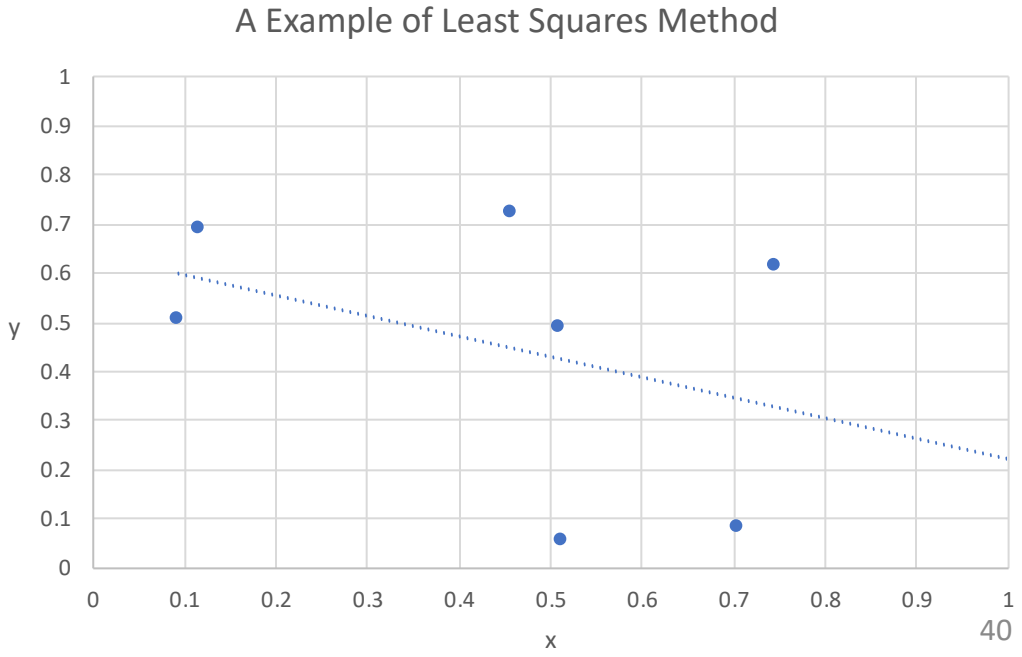
Regularized Regression Model

An analysis method that reduces the estimate by adding constraints to the ordinary least squares method

Features:
1) Calculation of estimation
2) Variable selection

Examples:
1) Ridge regression
2) Lasso regression
3) Elastic Net regression

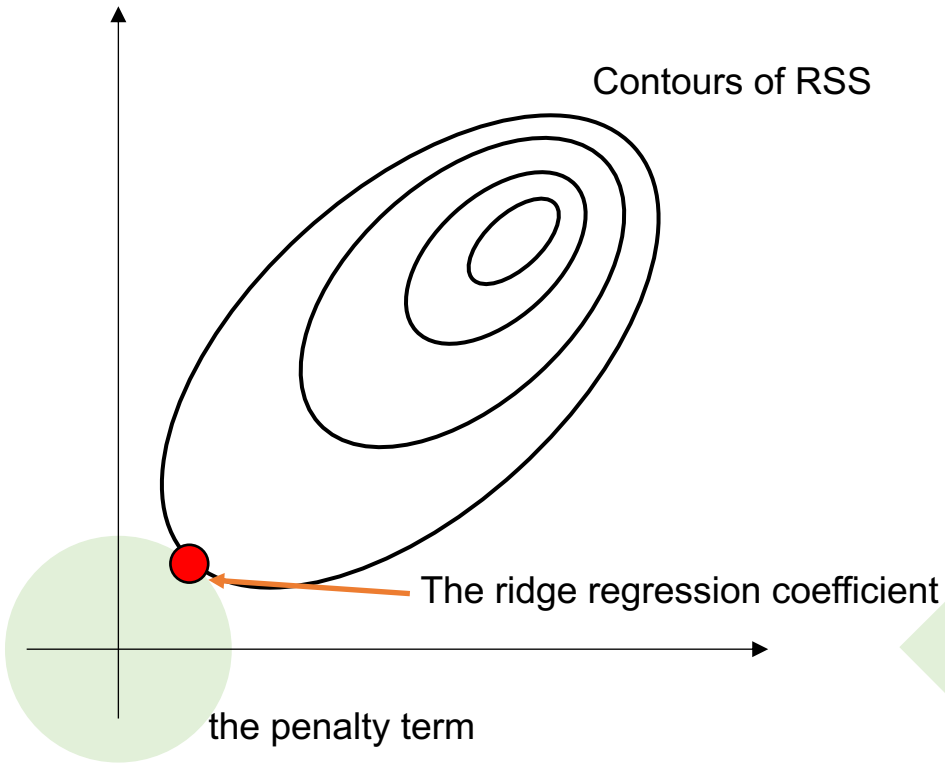
x	y
0.85823458	0.0172966
0.08798015	0.79650097
0.32214503	0.97746291
0.60233696	0.47201085
0.34828249	0.61622624
0.11730202	0.22664236
0.47086469	0.10788457



Ridge Regression and Lasso Regression

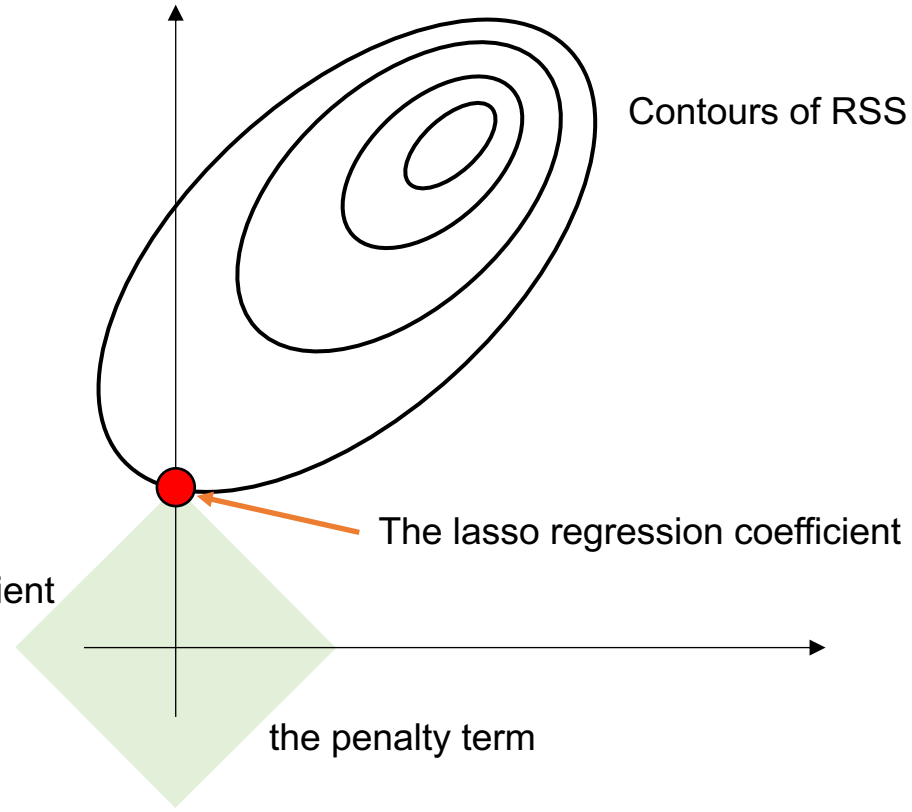
Ridge regression

It is used to improve the accuracy of the model.



Lasso regression

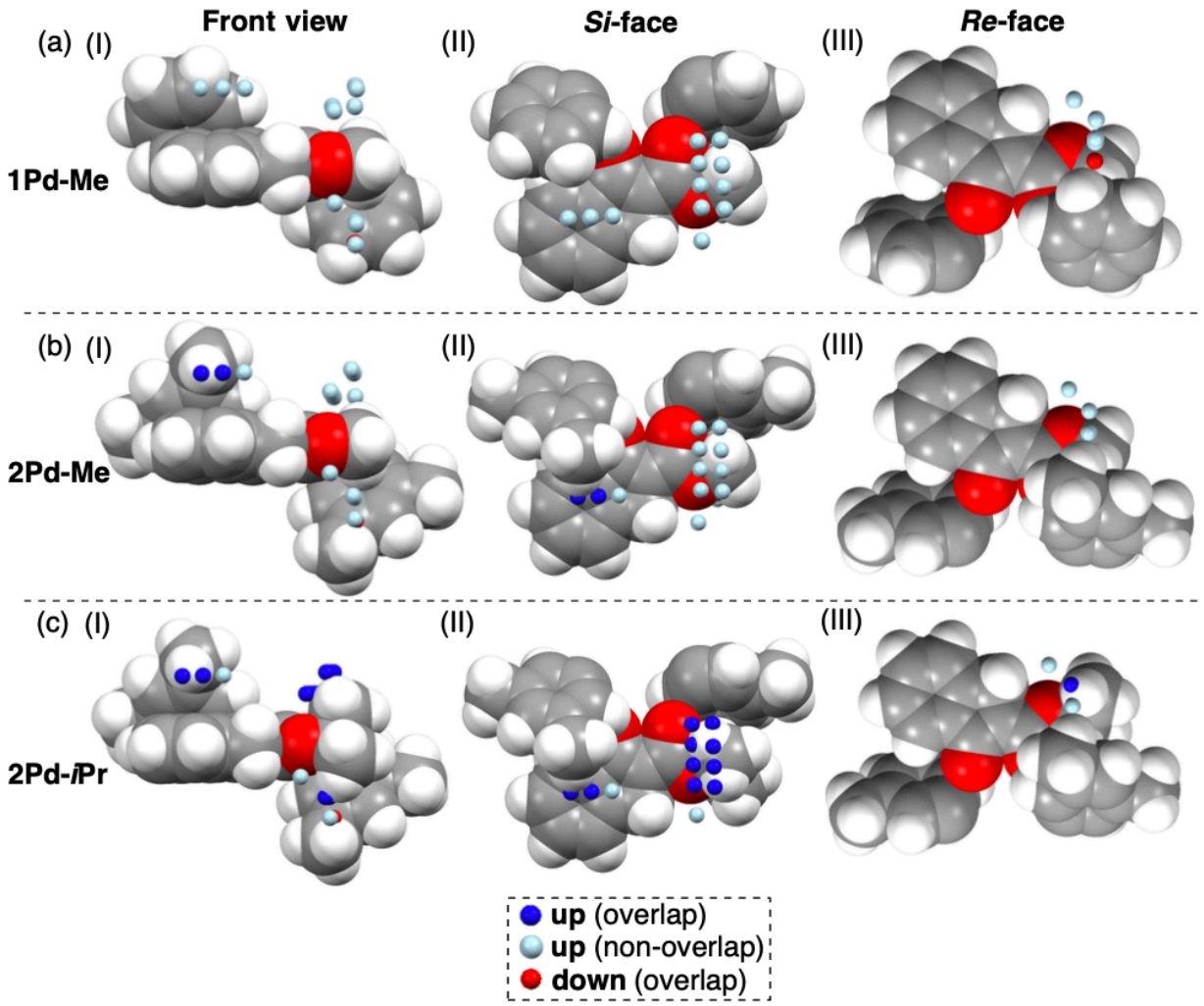
The number of variables included in the model is limited, making it easier to interpret.



RSS: Residual Sum of Squares

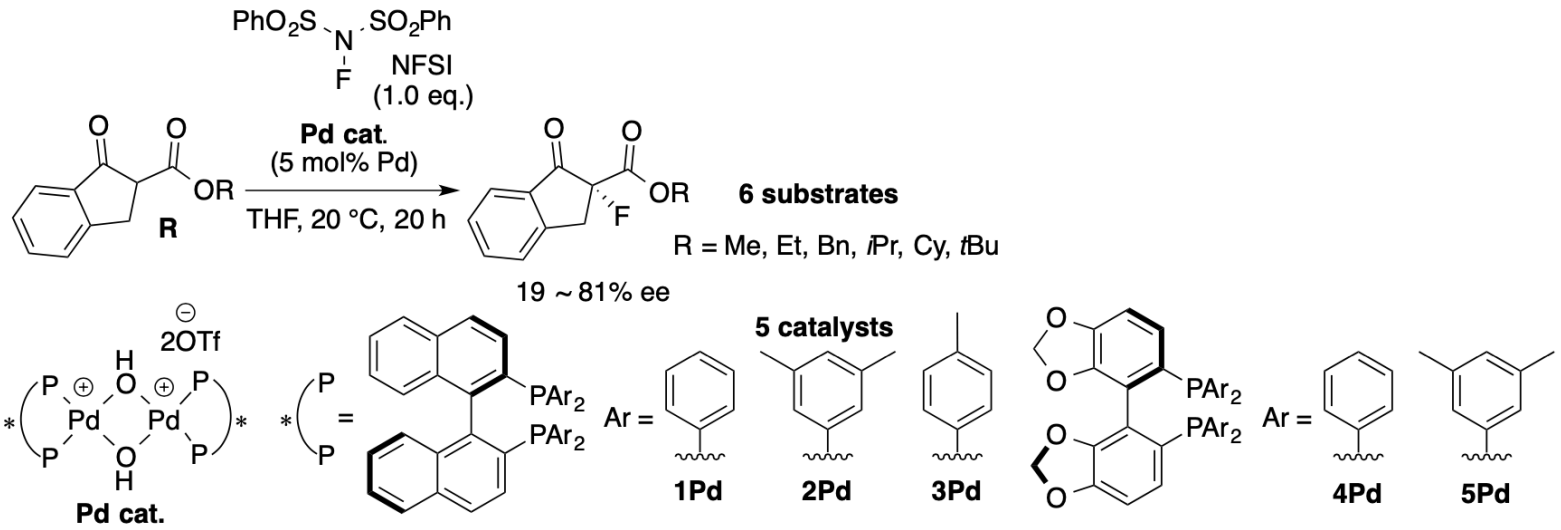
Elastic Net has the strong points of both the ridge regression and the lasso regression.

Visualized important structural information with the structures



1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

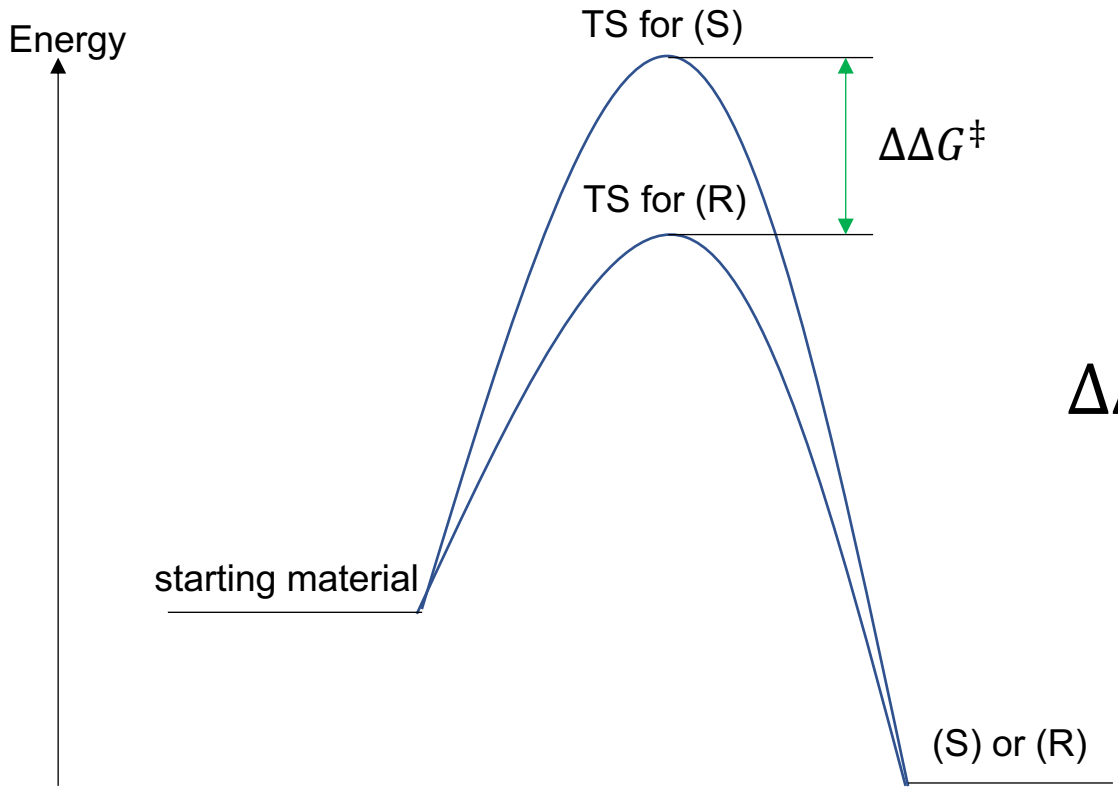
Data Used for the MFA



Catalyst/Substrate	Me	Et	Bn	*i*Pr	Cy	*t*Bu	Bzh
1Pd	36 (0.44)	28 (0.34)	40 (0.49)	38 (0.47)	46 (0.59)	52 (0.68)	77 (1.19)
2Pd	60 (0.81)	63 (0.86)	69 (0.99)	77 (1.19)	77 (1.19)	81 (1.32)	94 (2.0)
3Pd	38 (0.47)	30 (0.36)	40 (0.49)	48 (0.61)	55 (0.72)	61 (0.83)	83 (1.39)
4Pd	29 (0.34)	19 (0.22)	33 (0.40)	29 (0.35)	37 (0.46)	37 (0.45)	75 (1.13)
5Pd	60 (0.8)	63 (0.86)	70 (1.0)	78 (1.23)	79 (1.24)	81 (1.32)	97 (2.44)

1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Arrhenius equation



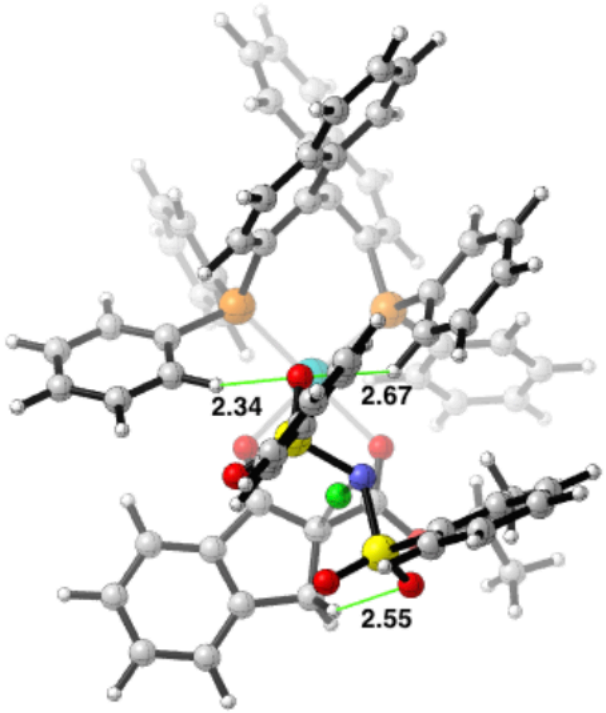
$$\Delta\Delta G^{\ddagger} = -RTlnk_{rel}$$

$\Delta\Delta G^{\ddagger}$: a difference in free energy between competing transition states leading to different enantiomers

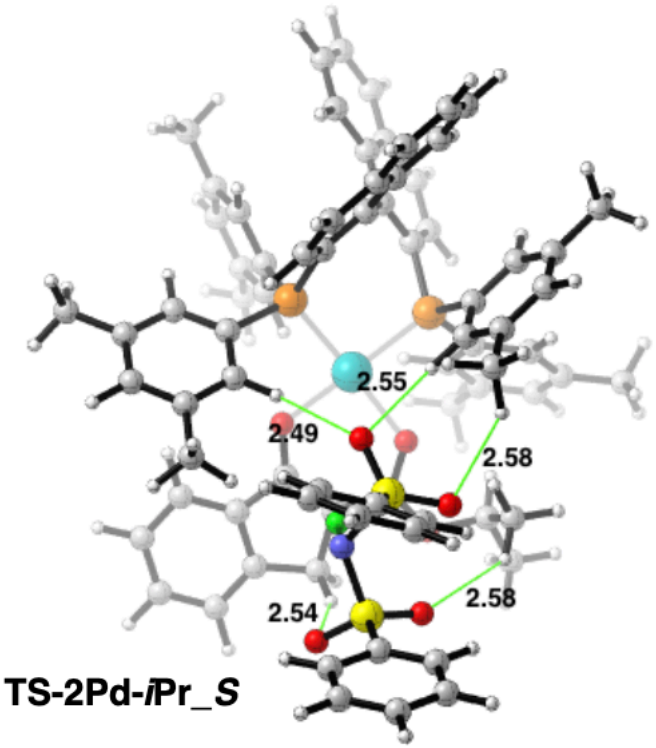
$k_{rel} = \frac{100+\%ee}{100-\%ee}$, R: gas constant, T: temperature

1) Toyao,T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. *ACS. Catal.* **2020**, *10*, 2260

NFSI conformation of the transition states

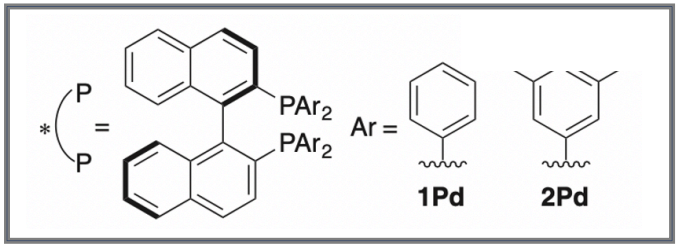


TS-1Pd-*i*Pr_*S*



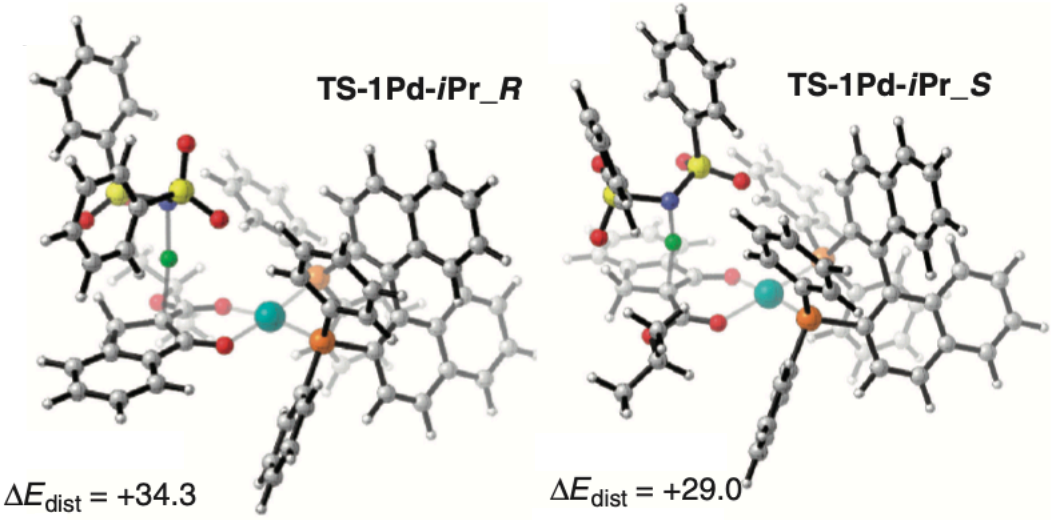
TS-2Pd-*i*Pr_*S*

The lowest energy conformers are showed in this page. Distances between oxygen atoms on NFSI and some hydrogen atoms on the Pd-enolate complexes in the structures The units of atomic distances are Å.
NFSI conformations are different.

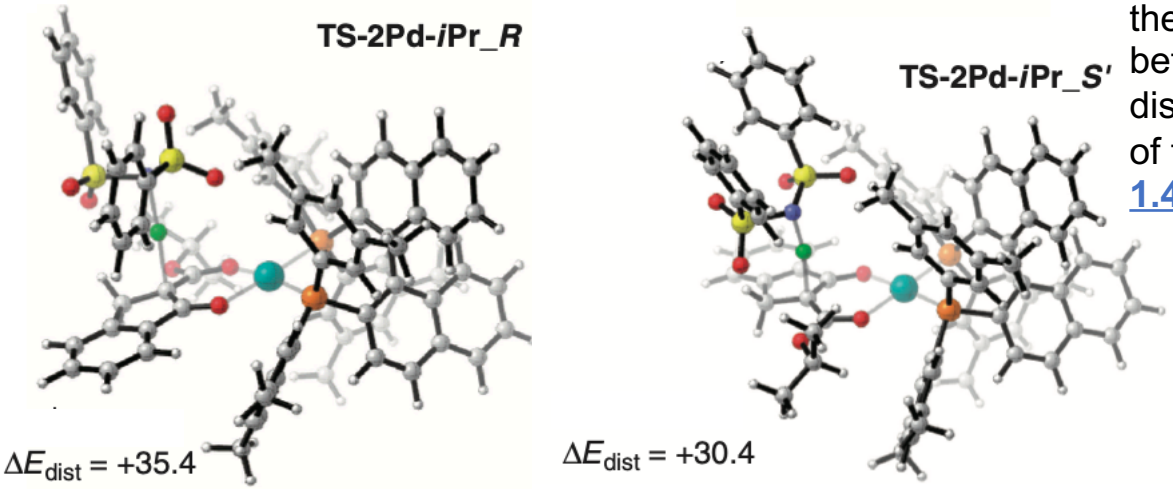


1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Distortion Energy



the second lowest energy geometry in S-pathway of the reaction 2Pd-iPr had the same NFSI conformation with that of TS-1Pd-iPr_S

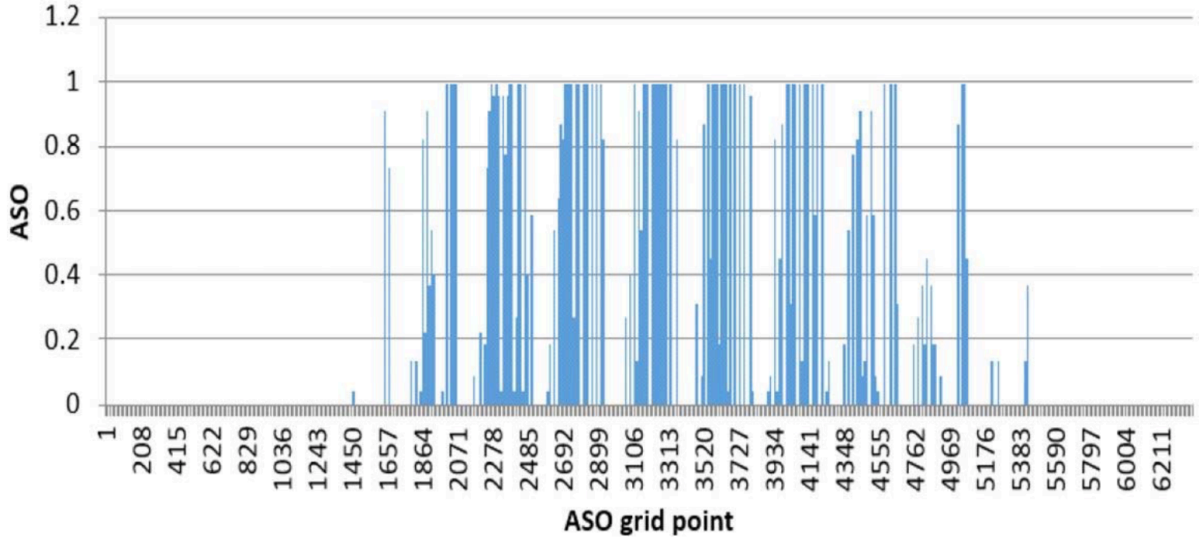


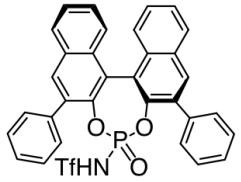
the difference between the distortion energies of the S-pathways is **1.4 kcal/mol**

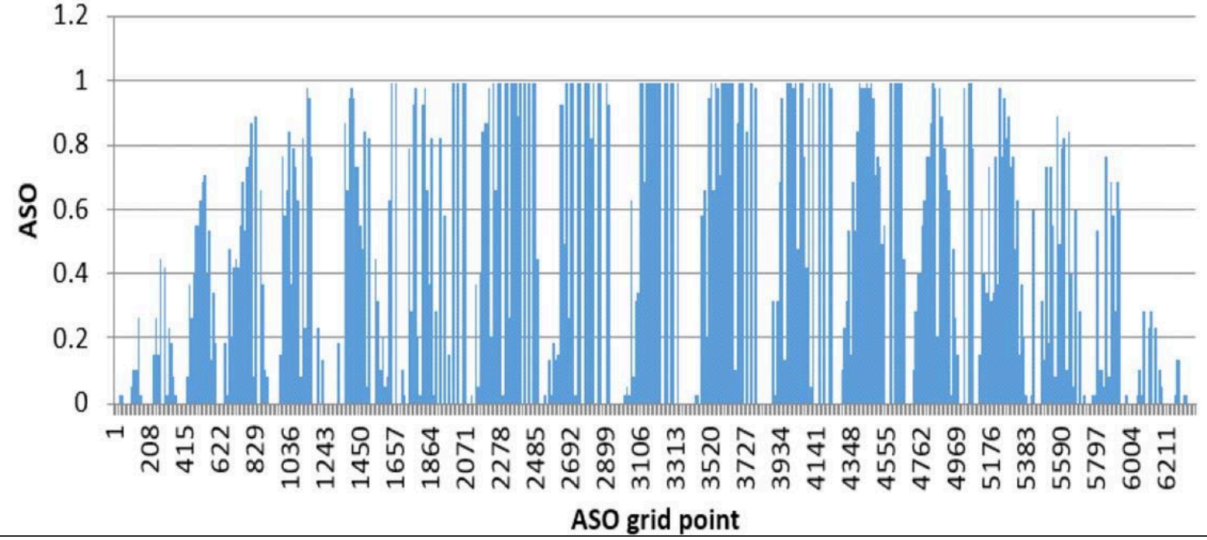
the difference between the distortion energies of the R-pathways is **1.1 kcal/mol** c

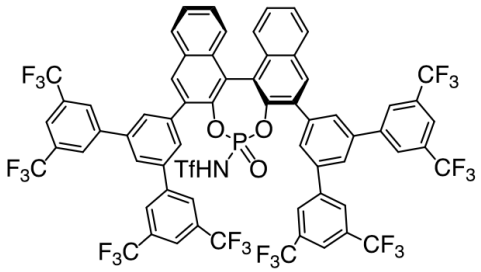
1) Yamaguchi, S.; Sodeoka, M. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1701

Average Steric Occupancy (ASO) (step 2)









1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Electrostatic Potential (ESP) (step 2)

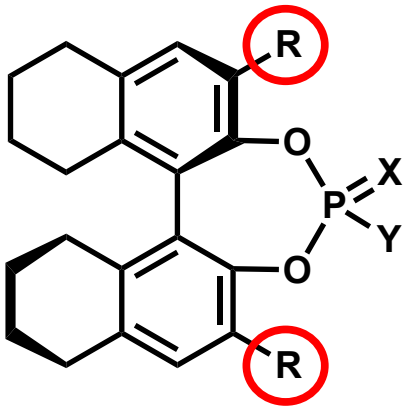
Conventional methods (Hammett parameters)

- represent the through-bond electronic perturbation a substituent has on a system

- However, the 3,3′- substituents in the *in silico* library are too diverse to be represented with experimentally derived Hammett parameters

TABLE I. Hammett and Modified Swain–Lupton Constants[a,d]

	substituent	σ_m	σ_p
1.	BF_2	0.32	0.48
2.	Br	0.39	0.23
3.	$GeBr_3$	0.66	0.73
4.	$SiBr_3$	0.48	0.57
5.	Cl	0.37	0.23
6.	HgCl	0.33	0.35
7.	SO_2Cl	1.20	1.11
8.	SCl	0.44	0.48
9.	ICl_2	1.10	1.11

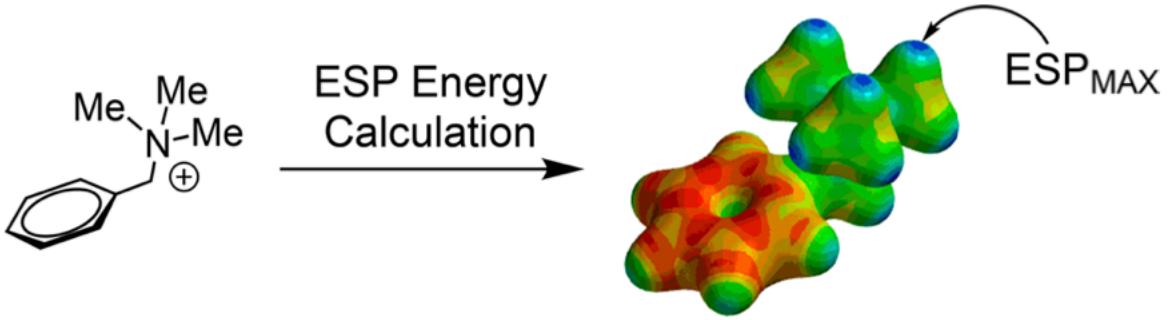


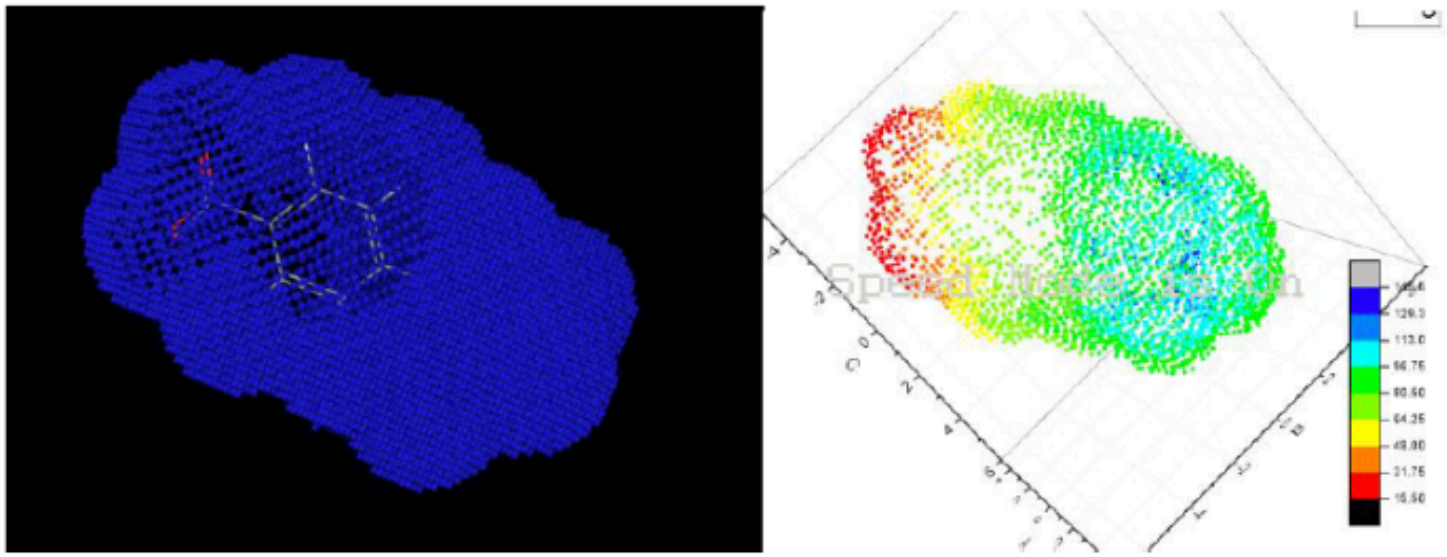
3,3'- substituents

→ a new calculable parameter had to be developed that reflects the perturbation of the substituent on a charged particle: ESP

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631
2) Henle, J. J.; Zahrt, A. E.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. *J. Am. Chem. Soc.* **2020**, *142,* 11578
3) Hansch, C.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991**, *91,* 165

Electrostatic Potential (ESP) (step 2)

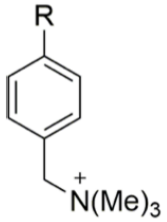




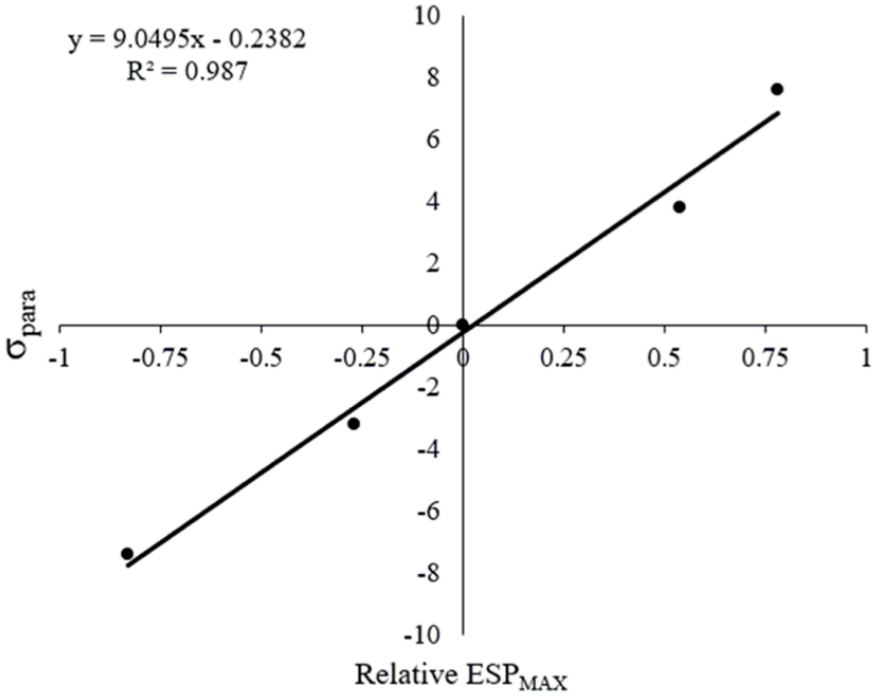
Example MIF calculated for 4-nitrobenzyltrimethylammonium cation.

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631
2) Henle, J. J.; Zahrt, A. E.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. *J. Am. Chem. Soc.* **2020**, *142,* 11578

Electrostatic Potential (ESP) (step 2)



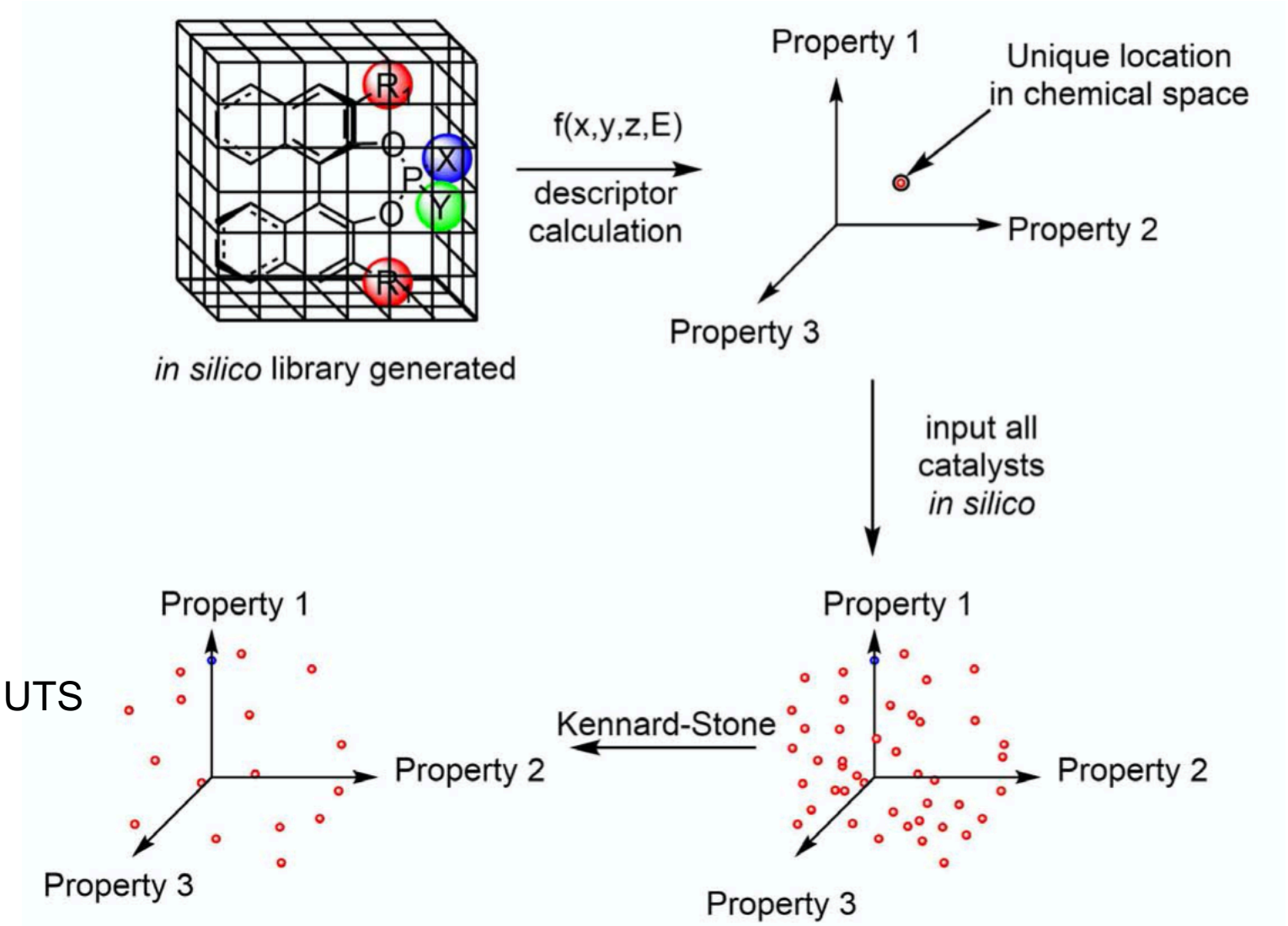
R	ESP_{MAX} (kcal/mol)	Rel. ESP_{MAX}	σ_{para}
NO_2	131.5	+7.6	0.78
CF_3	128.0	+3.8	0.54
H	124.2	0	0
OMe	121.0	-3.2	-0.27
NMe_2	116.8	-7.4	-0.83



Evaluation of ESP_{MAX} descriptor by correlating relative ESP_{MAX} with Hammett parameters.

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631
2) Henle, J. J.; Zahrt, A. E.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. *J. Am. Chem. Soc.* **2020**, *142,* 11578

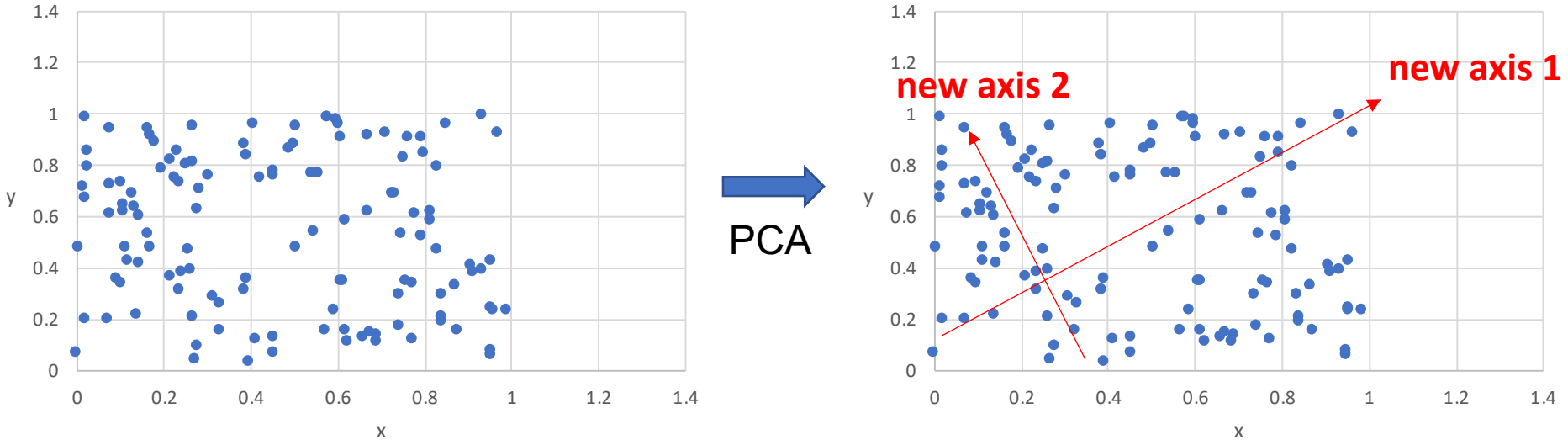
Universal Training Set (UTS) (step 3)



1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Principal Component Analysis (PCA)

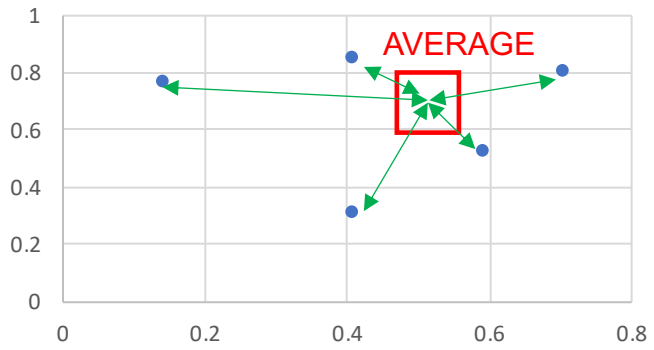
Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.

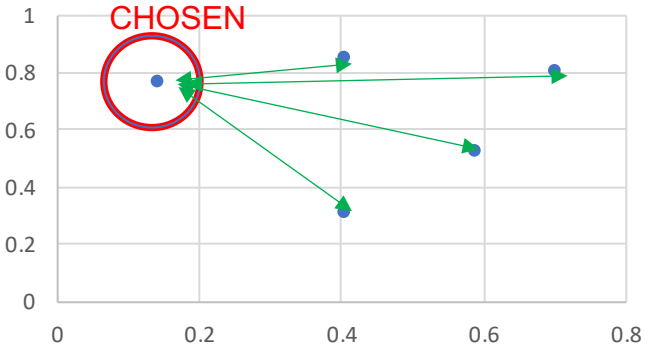


1) Jolliffe, I. T.; Cadima, J. *Phil. Trans. R. Soc. A.* **374**: 20150202

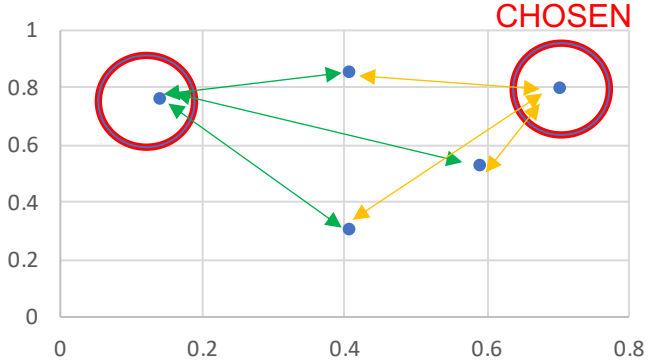
Kennard-Stone Algorithm (step 3)

This algorithm is one of the famous ways to select a sample of training data in an evenly distributed manner





Chose the maximum of the representative distance

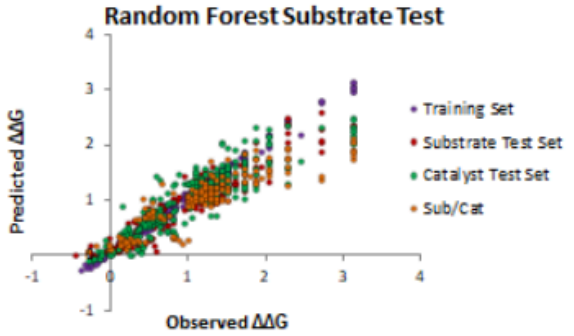


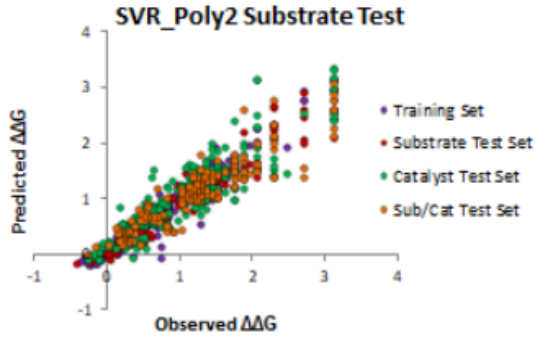
repeat

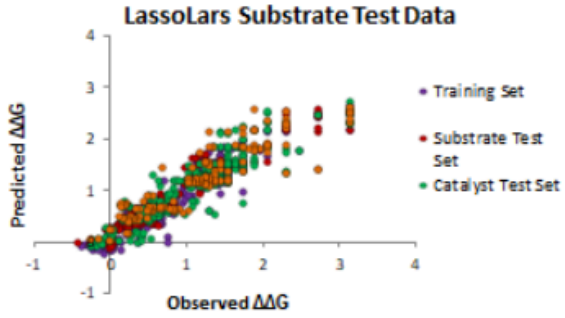
Calculate Euclidean distances between the not yet selected samples and all the samples selected so far. The minima of those distances become the representative distance for each sample.

1) Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137

Machine Learning Method (step 4)





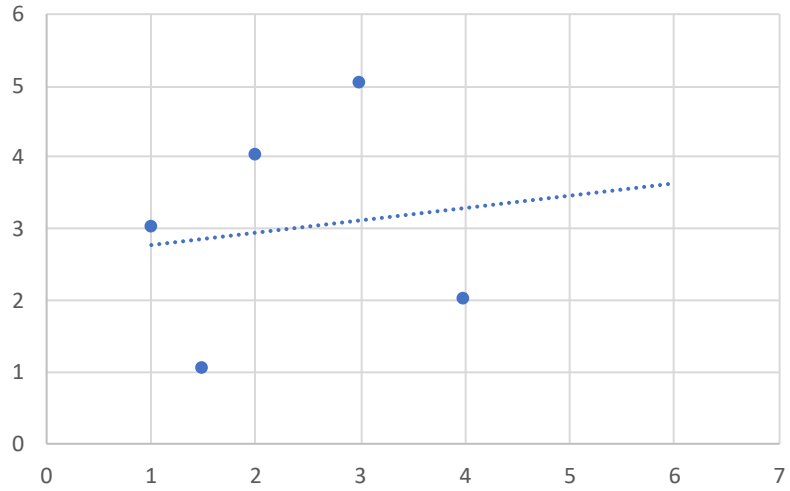


Entry	Model	Q^2 (k=5)	R^2	MAD (Train, ΔΔG)	MAD (SubTest, ΔΔG)	MAD (CatTest, ΔΔG)	MAD (Sub/CatTest, ΔΔG)
1	RF	0.731	0.991	0.048	0.211	0.211	**0.294**
2	SVR_POLY2	0.748	0.953	0.096	0.161	0.211	**0.238**
3	LASSOLARS	0.572	0.906	0.160	0.165	0.206	**0.224**

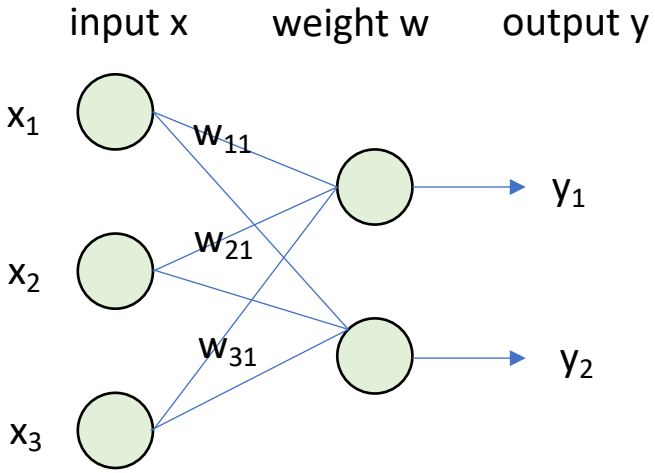
1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Parametric method (step 4)

Linear Regression



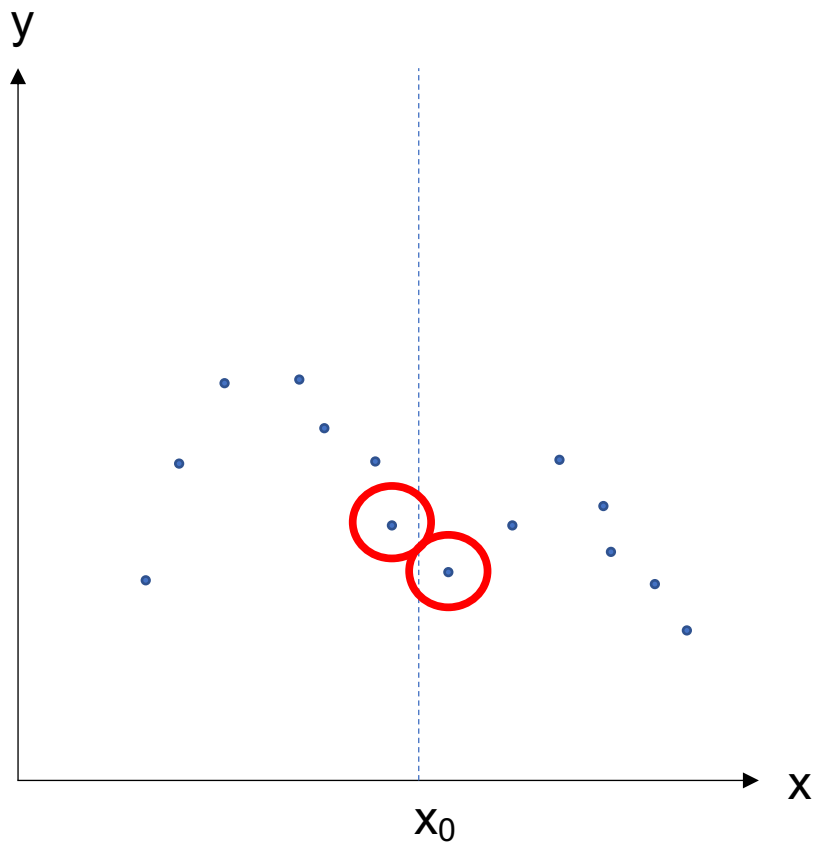
Feedforward Neural Network



$y_1 = f(w_{11}x_1 + w_{12}x_2 + w_{13}x_3)$

These parameters are determined by training sets. The training sets are never used after training.

Kernel method (step 4)



Predict the value of y when $x=x_0$

It is thought to take a value close to y of the data (○) in the vicinity of x_0

On the other hand, there is no relationship between the data far from x_0 and the value of y when $x=x_0$.

$$y_0 = \sum_{k=0}^{n} (Function\ that\ represents\ the\ distance\ between\ x_0\ and\ x_k) \times y_k$$

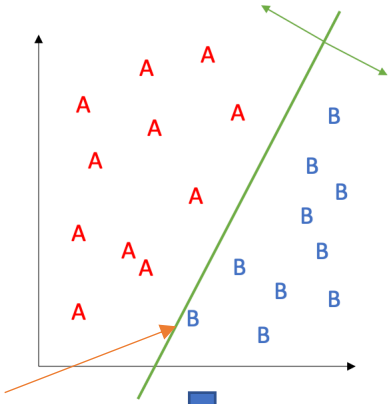
Kernel function

Support Vector Machine (SVM) (step 4)

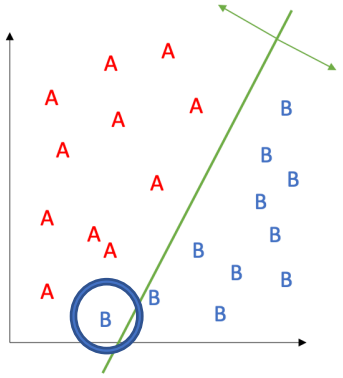
It is thought that SVM is has the best pattern classification in the common machine learning methods.

conventional method

classification



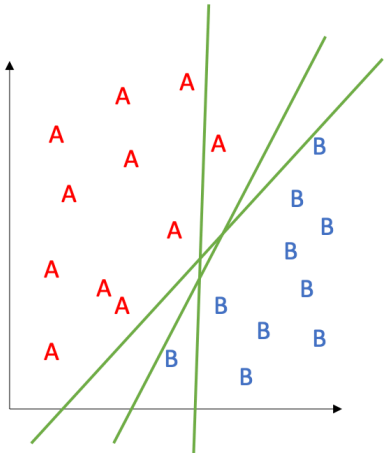
near the boundary line

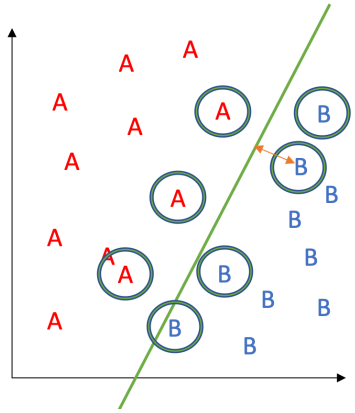


This model tends to make a wrong prediction

SVM

identify the ”best boundary line”



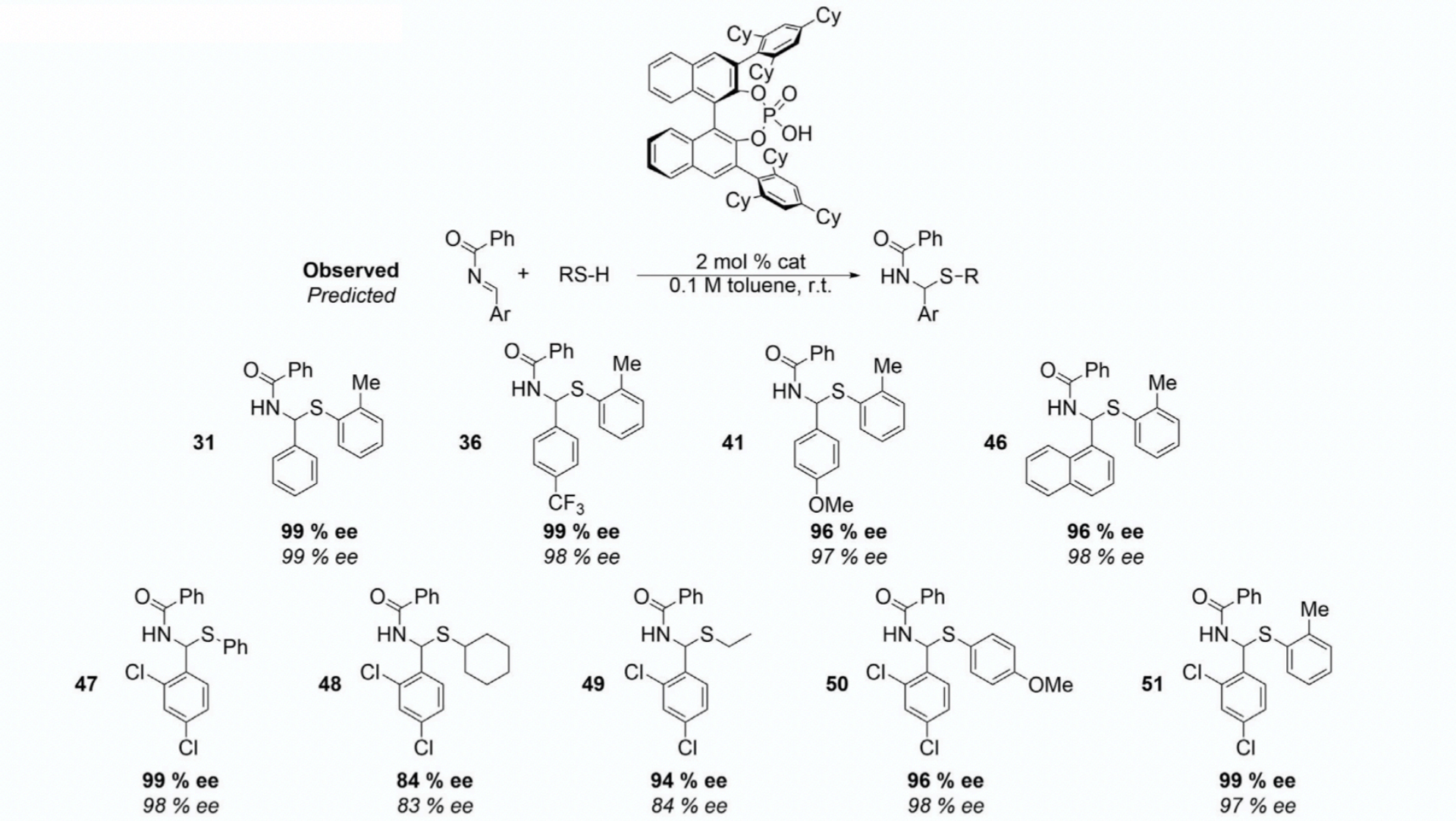


：support vectors

：margin

“margin” is a distance between a support vector and a boundary line

Examples of Good Prediction

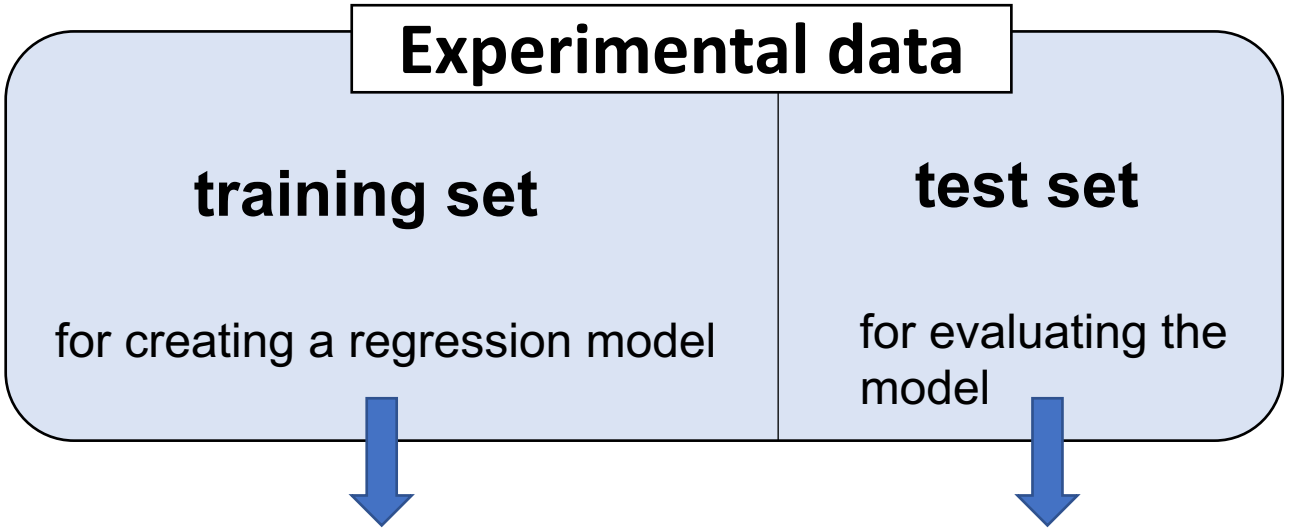


1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Another Modeling Study

Could this tool be used to identify new reactions that are more selective than any reaction in the training data?

In this model study…

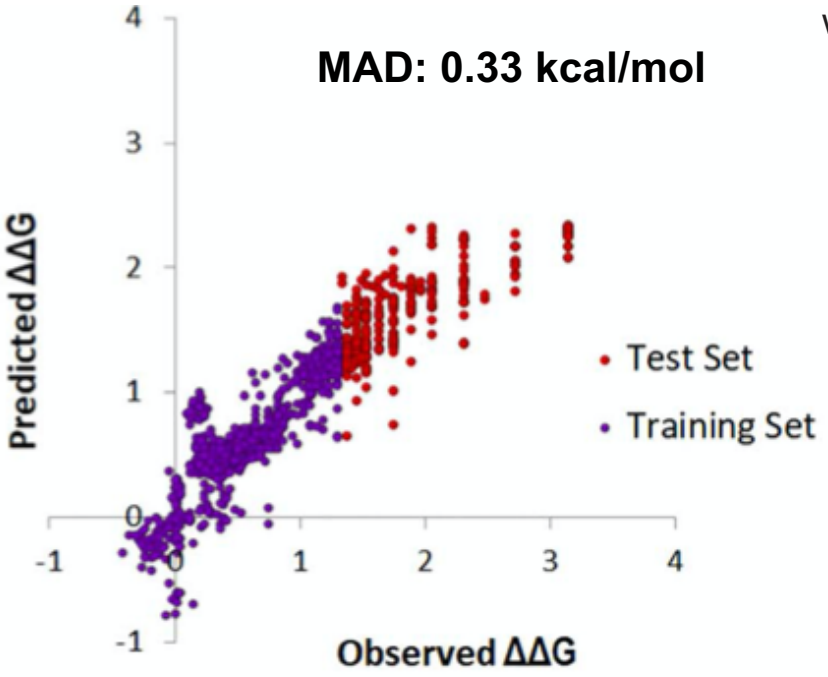


Note that the UTS was not used in this model study.

1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

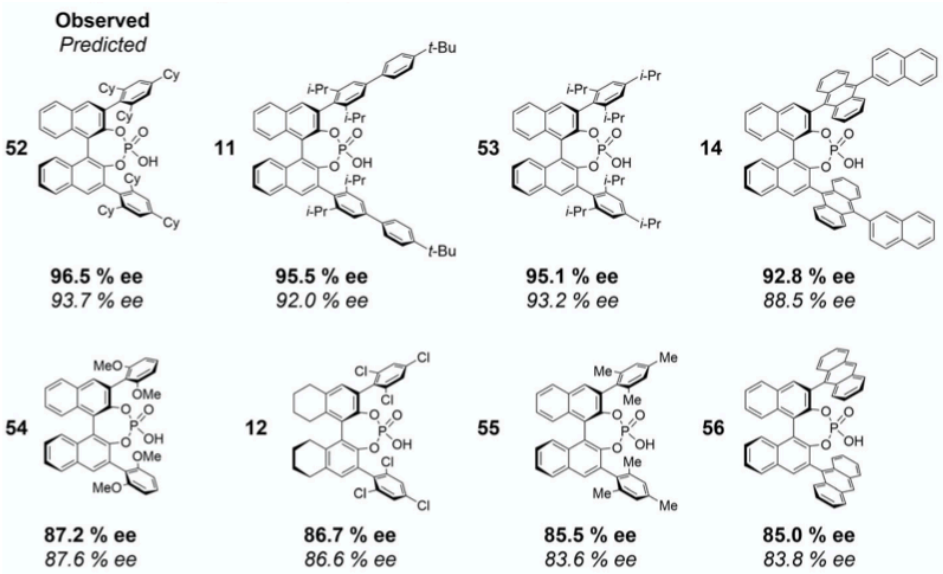
Another Modeling Study

Could this tool be used to identify new reactions that are more selective than any reaction in the training data?



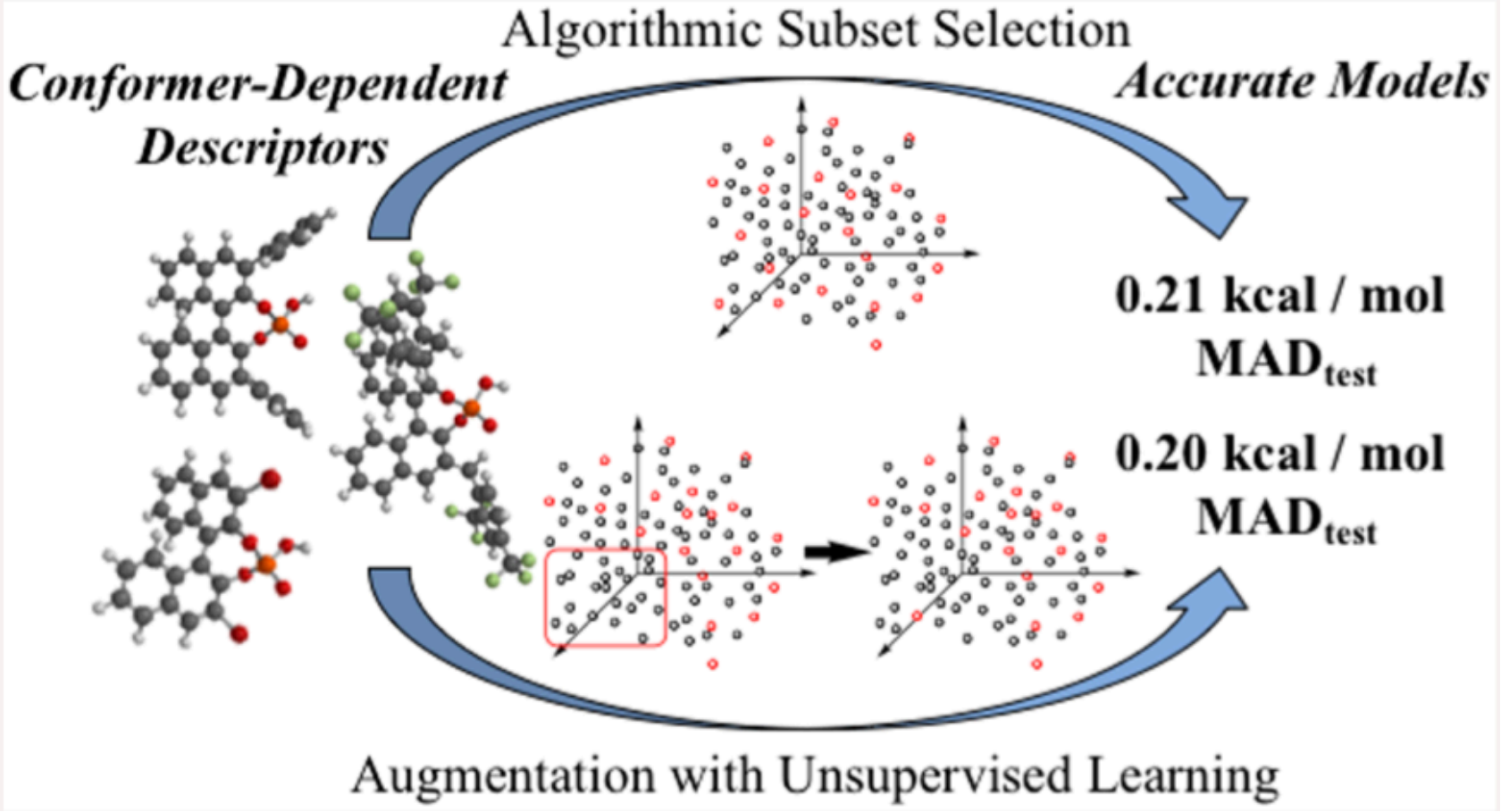
(deep feed-forward neural networks)

Predicted and observed average selectivities for the 8 catalysts with average enantioselectivity over 80% ee are shown below.



1) Zahrt, A. E.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*, eaau5631

Development of a Computer Workflow for Catalyst Optimization



1) Henle, J. J.; Zahrt, A. E.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. *J. Am. Chem. Soc.* **2020**, *142,* 11578