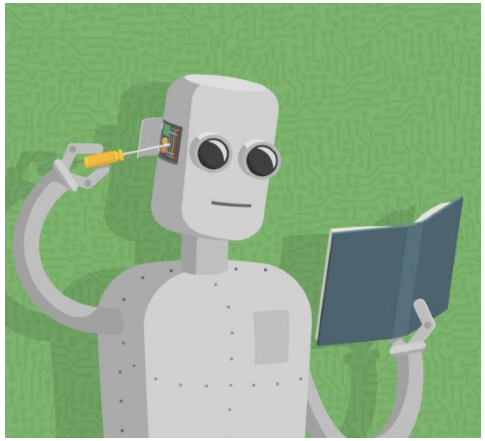
Predicting Reaction Performance Using Machine Learning

2018.4.21. Literature Seminar
M2 Yusuke Imamura



Contents

1. Introduction

2. Prediction in Intermolecular Reaction Screening (main paper)

3. Prediction in Usual Reaction Screening

Machine Learning (ML)

（前略）…先日、あの羽生善治永世七冠に直接お会いする機会があり、AIに関する非常に興味深い話を伺いました。羽生永世七冠は、ご自身がAI技術に対して非常に大きな興味を持っており、将棋を通じてAI技術も研究しているそうです。AIがその膨大なデータ処理能力を通じて、人間が長年かけて築き上げてきた将棋の世界の境界をときに飛び越え、これまでとは違う将棋の姿を見せてくれるというのです。永世七冠の羽生さんの、新しい可能性と向かい合い、面白いと思って挑戦し続ける姿勢は、ぜひ、参考にしてもらいたいと思います。…（後略） 平成29年度東京大学学位記授与式　総長告辞

「AIが仕事を奪う」　脅威論への素朴な疑問

B!　PUSH通知

4月13日（金）9時0分　ITmedia NEWS

「人工知能（AI）が原因で失業する」と信じている人は大勢います。では私たちの周りに、人工知能に仕事を奪われた人は居るでしょうか？　少なくとも、今の所は私の周りにはいません。もしかして私たちは、居るはずのない幽霊にただおびえているのではないでしょうか。

そのような怪談が語られ始めたのは、第3次人工知能ます。オズボEMPLOJOBS TO COMPUTERISATION？」という論文47%が機械に代替されるリスクがある」というに花開きました。たった20年で労働者の半数が失然とするのも当然です。

Why ML is hot area now?
- ✔ **Improvement in calculation**
- ✔ **Availability of big data**
- ✔ **Information sharing**

シマンテック、EDR製品で標的型攻撃を検知する機械学習活用機能

4/18(水) 7:00配信

ASCII

シマンテックが、同社のEDR製品「Symantec Advanced Threat Protection（ATP）」において、機械学習を適用して標的型攻撃の検出などを行う新機能の追加を発表した。

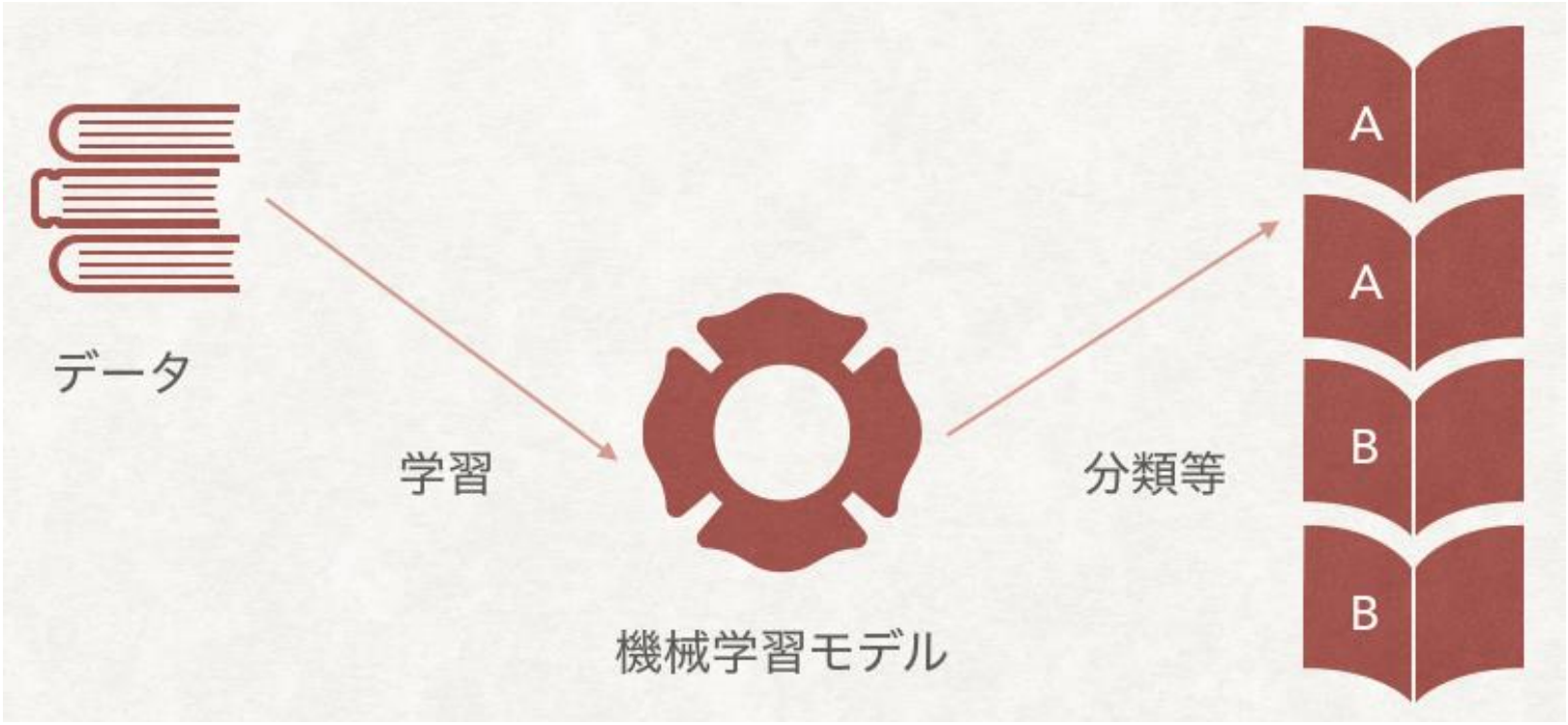
1) https://www.u-tokyo.ac.jp/gen01/b_message29_09_j.html
2) http://www.itmedia.co.jp/news/articles/1804/13/news018.html 3) http://ascii.jp/elem/000/001/665/1665363/

What is ML?

Machine learning is one of research topic in Artificial Intelligence (AI).

Machine (computer) learns pattern from large amount data, and then make a model for classification or regression (回帰).

cf. AI…Intelligence demonstrated by machine



Application in Daily Life

Recommendation in Amazon

Siri

Self driving technology

Application in Chemistry

• **Biomedical science**

Virtual screening of libraries of drug-like molecules for biological function.

- **See also 160723_LS_Shunichiroh_KATOH**

• **Organic chemistry**

Assist with synthetic planning via retrosynthetic pathway.

- **See also 160916_LS_Yuki_NAKAGAWA**

Predict the product(s) of chemical reactions given a set of reactants and conditions. (Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. *ACS Cent. Sci.* 2016, *2*, 725.)

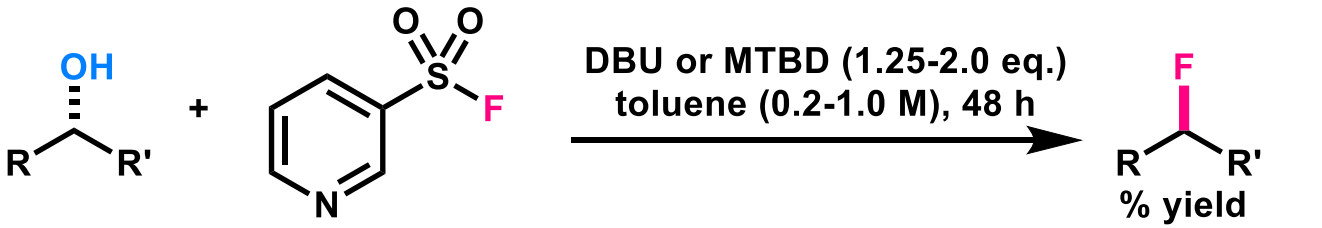
Predict the products of textbook questions.

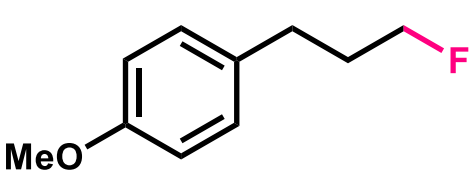
Challenges… It is difficult to use for non-specialists

Multidimensionality of chemical structure

Predicting Reactions

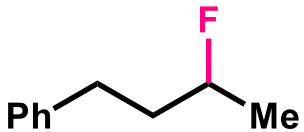
- Predict from substrate scope



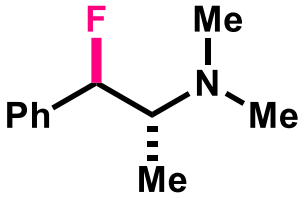


79%

primary and secondary alcohols work well

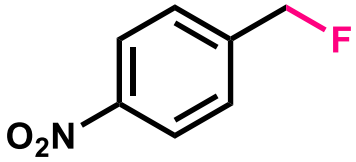


79%



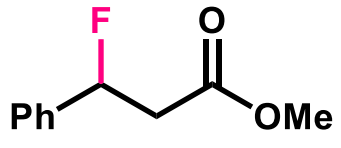
68% (>20:1)

compatible with unprotected amine



43%

not good for unhindered benzylic alcohol



-

(96%)

E1cb elimination proceed

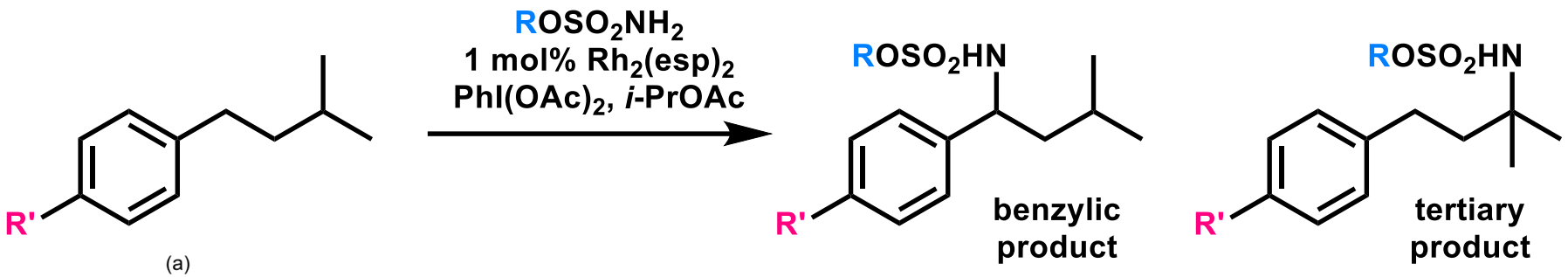
Chemical theory to understand the rule is needed.

1) Nielsen, M. L.; Ugaz, C. R.; Li, W.; Doyle, A. G. *J. Am. Chem. Soc.* **2015**, *137*, 9571

Predicting Reactions

- **Linear regression model (線形回帰)**

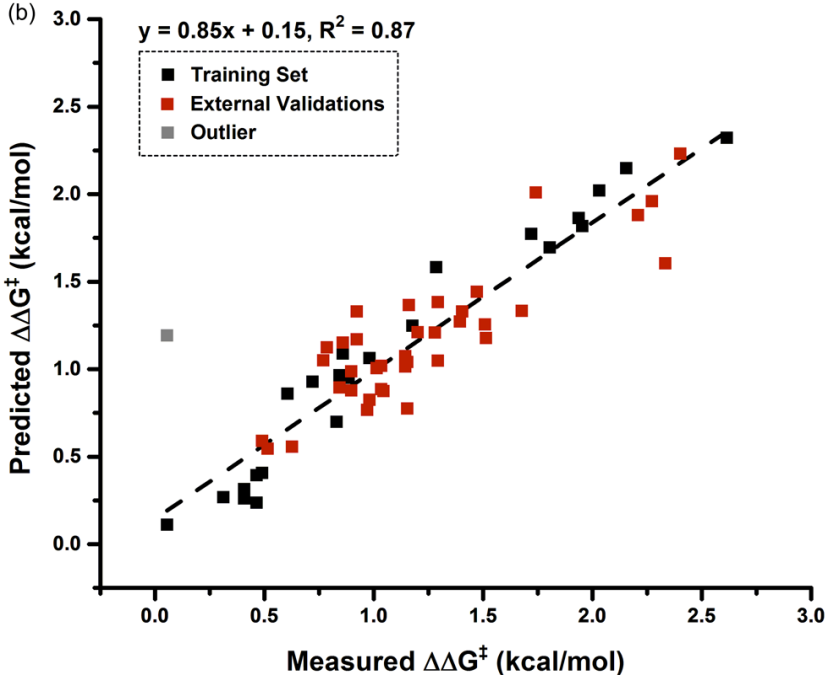
Site selective C-H amination (Bess, E. N.; Du Bois, J. *et al.*, *J. Am. Chem. Soc. 2014, 136, 5783)*



(a)

$-RT\ln(T/B) = \Delta\Delta G^{\ddagger} = -0.01 - 0.86(\sigma^{+}) - 0.11(\nu_{CO})(\sigma^{+}) + 0.17(\nu_{CO}) + 0.26(I_{OSN})$

(b)



selectivity correlates to
- **Hammet σ^{+} on R'**
- **two IR measures on R**

Difficulties:
The users must hand-select parameters according to known experience or theory.

Prof. Abigail G. Doyle

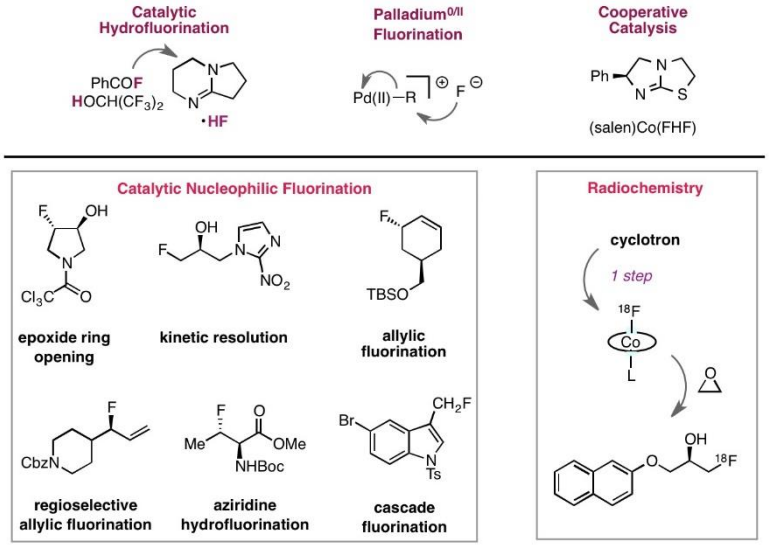
1998-2002 A.B. and A.M. @Harvard Univ. (Prof. Eric Jacobsen)
Research topic: iron-catalyzed epoxidations of alkenes

2002-2003 Pre-D. @Stanford Univ. (Prof. Justin Du Bois)
Research topic: gold catalysts for the hydration of unactivated alkenes

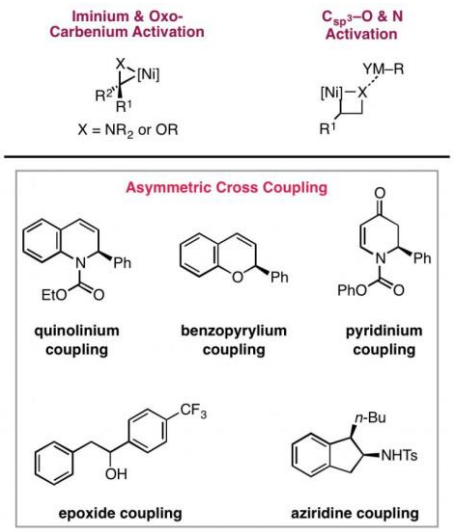
2003-2008 Ph.D. @Harvard Univ. (Prof. Eric Jacobsen)
Ph.D. Thesis: Engaging Alkyl Halides and Oxocarbenium Ions in Asymmetric Catalysis.

2008-2013 Assist. Prof. @Princeton Univ. Department of Chemistry
2013-2017 Associate Prof.
2017- Professor

Nucleophilic Fluorination



Ni-catalyzed Cross Coupling



See also 171118_LS_Daiki_Kamakura

Contents

1. Introduction

2. **Prediction in Intermolecular Reaction Screening (main paper)**

3. Prediction in Usual Reaction Screening

REPORT

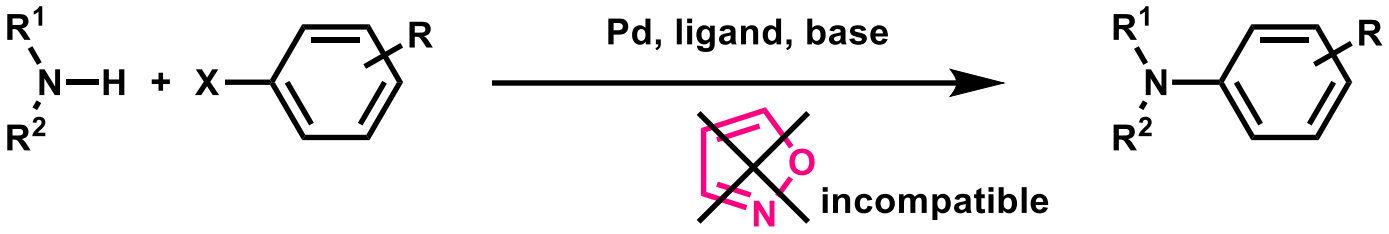
Predicting reaction performance in C–N cross-coupling using machine learning

Derek T. Ahneman[1], Jesús G. Estrada[1], Shishi Lin[2], Spencer D. Dreher[2,*], Abigail G. Doyle[1,*]

+ See all authors and affiliations

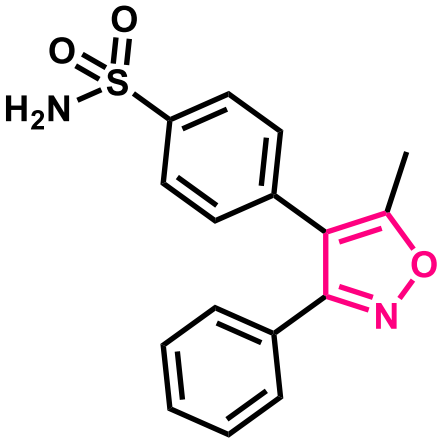
Science 13 Apr 2018:
Vol. 360, Issue 6385, pp. 186-190
DOI: 10.1126/science.aar5169

Buchwald-Hartwig Reaction [1)]

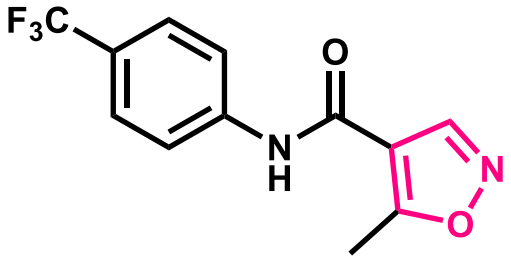


Useful reaction to synthesize pharmaceuticals.
But incopatible with heterocycles that contain heteroatom-heteroatom bonds (ex. isoxazole).

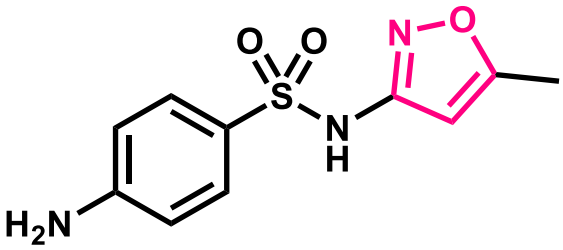
Pharmaceuticals containing an isoxazole core



Valdecoxib (Bextra)
COX-2 inhibitor



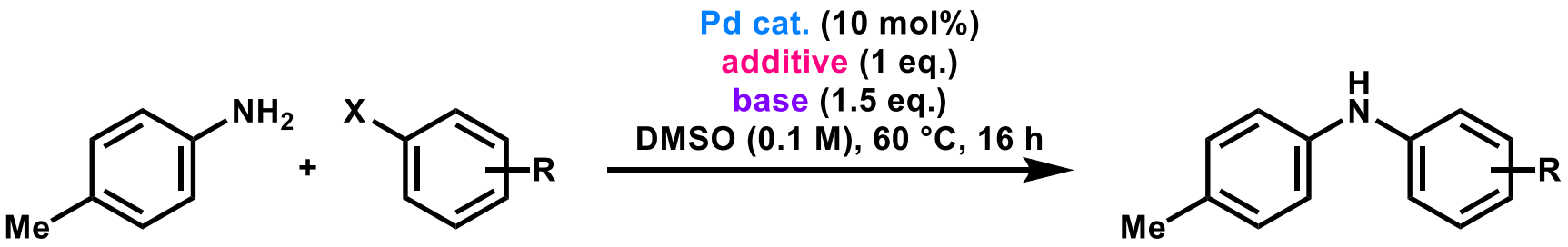
leflunomide (Arava)
DMARD

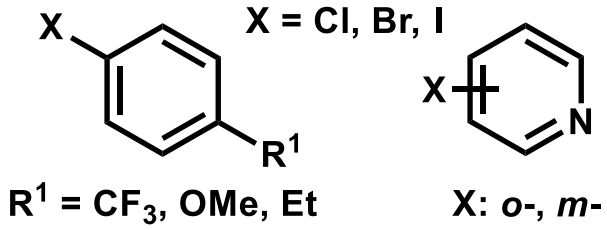


Sulfamethoxazole
antibiotics

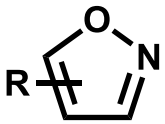
1) a) Paul, F.; Patt, J.; Hartwig, J. F. *J. Am. Chem. Soc.* **1994**. *116*, 5969.
b) Guram, A. S.; Buchwald, S. L. *J. Am. Chem. Soc.* **1994**. *116*, 7901.

Intermolecular Reaction Screening [1)]

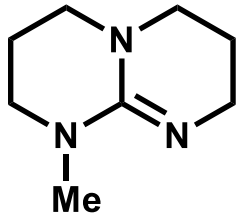




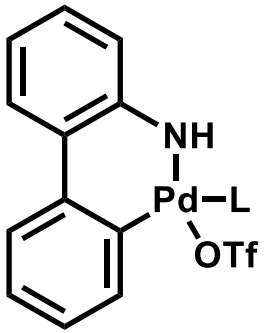
Aryl Halides×15

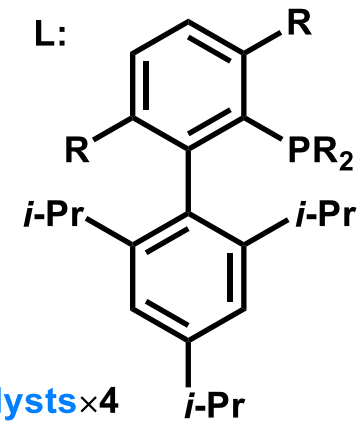


Additives×23

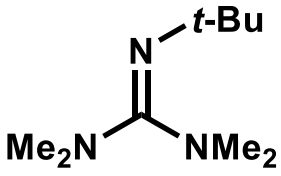


MTBD

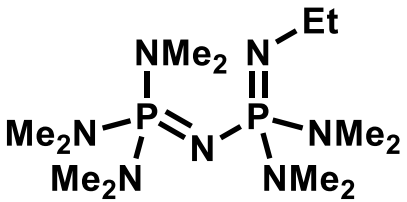




Pd Catalysts×4



BTMG



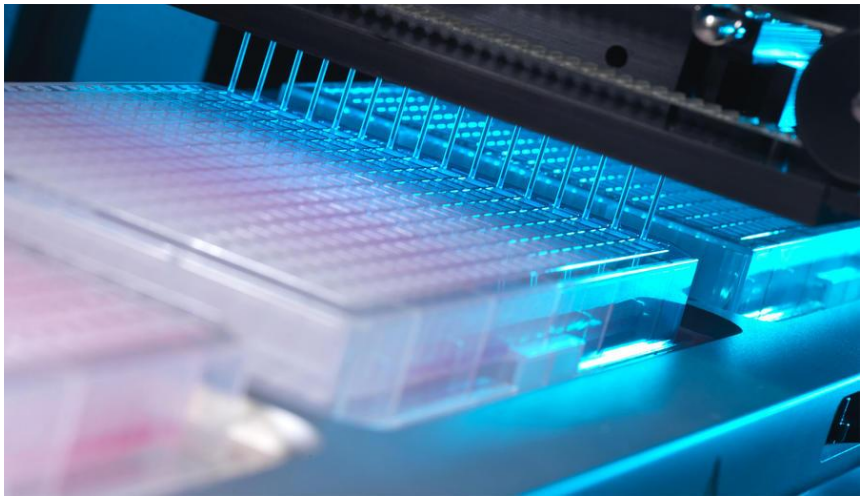
P2Et

Base×3

16×24×3×4 = 4608 reactions

1) Collins, K. D.; Glorius, F. *Nat. Chem.* **2013**, *5*, 597.

Ultra-High-Throughput Experimentation



Stock solutions in DMSO (200 nL each) was added with Mosquito HTS liquid handling robot.

1536-well plate×3

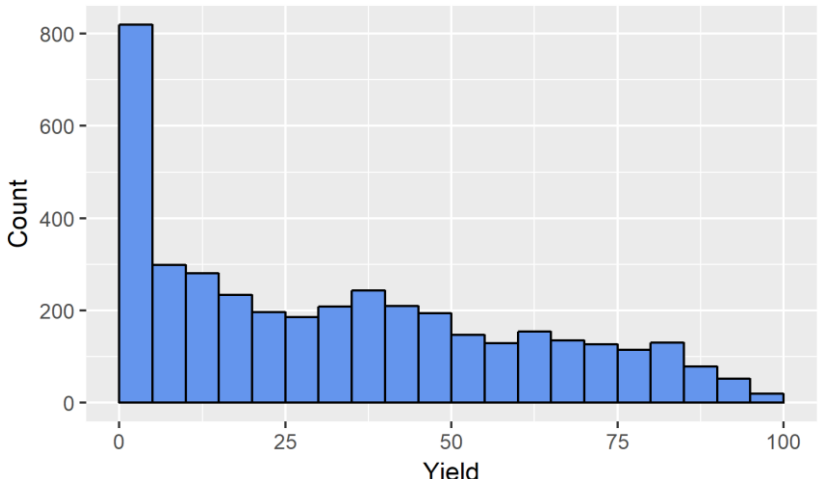
4608 reactions

catalyst, aryl halide
additive, base

internal standard
(di-*tert*-butylbiphenyl)

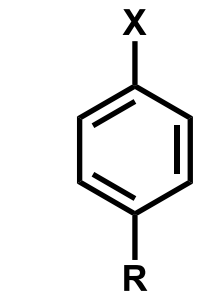
sealed, 60 °C, 16 h

UPLC analysis
UV210 nm detection

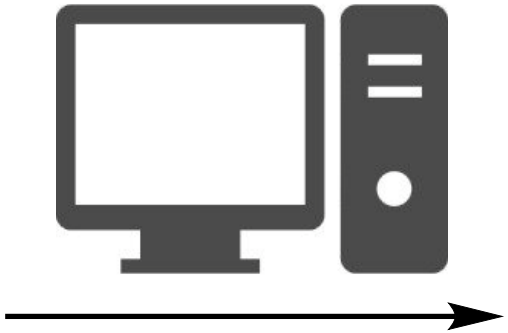


Descriptor

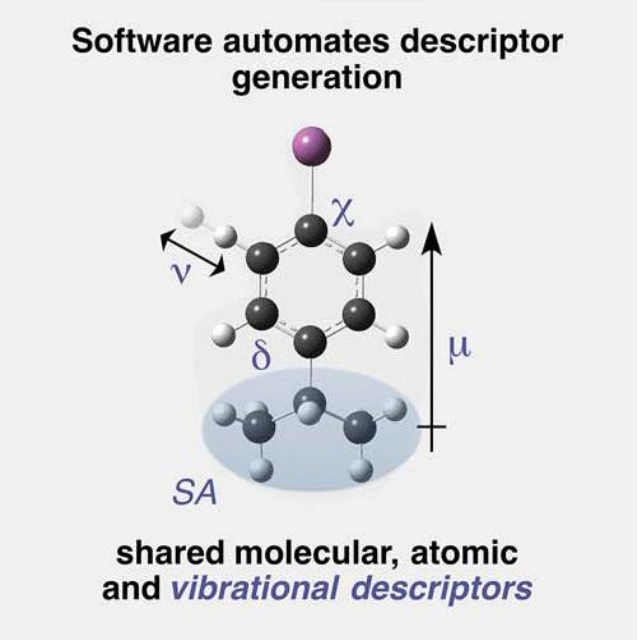
Descriptor: Variable that descripts chemical property of the substrate.



input structure



Calculation with Spartan B3LYP/6-31G*



- **Additive Descriptors (*n* = 19)**

(E_{HOMO}, E_{LUMO}, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular Weight, Ovality, Surface Area, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, *C4 Electrostatic Charge, *C5 NMR Shift, *C5 Electrostatic Charge, *N1 Electrostatic Charge, *O1 Electrostatic Charge, V1 Frequency, V1 Intensity)

- **Aryl Halide Descriptors (*n* = 27)**

- **Base Descriptors (*n* = 10)**

- **Ligand Descriptors (*n* = 64)**

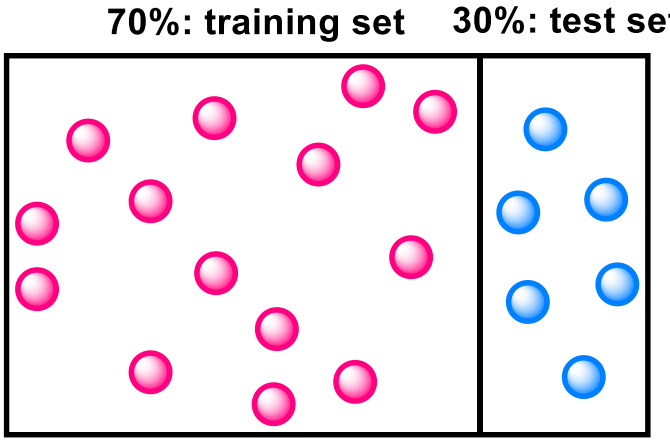
Prediction Procedure

total: 4608 reactions

70% → training set
(randomly chose from whole data)
relationship between yield and descriptors was evaluated by each method. (linear regression model and other ML methods)

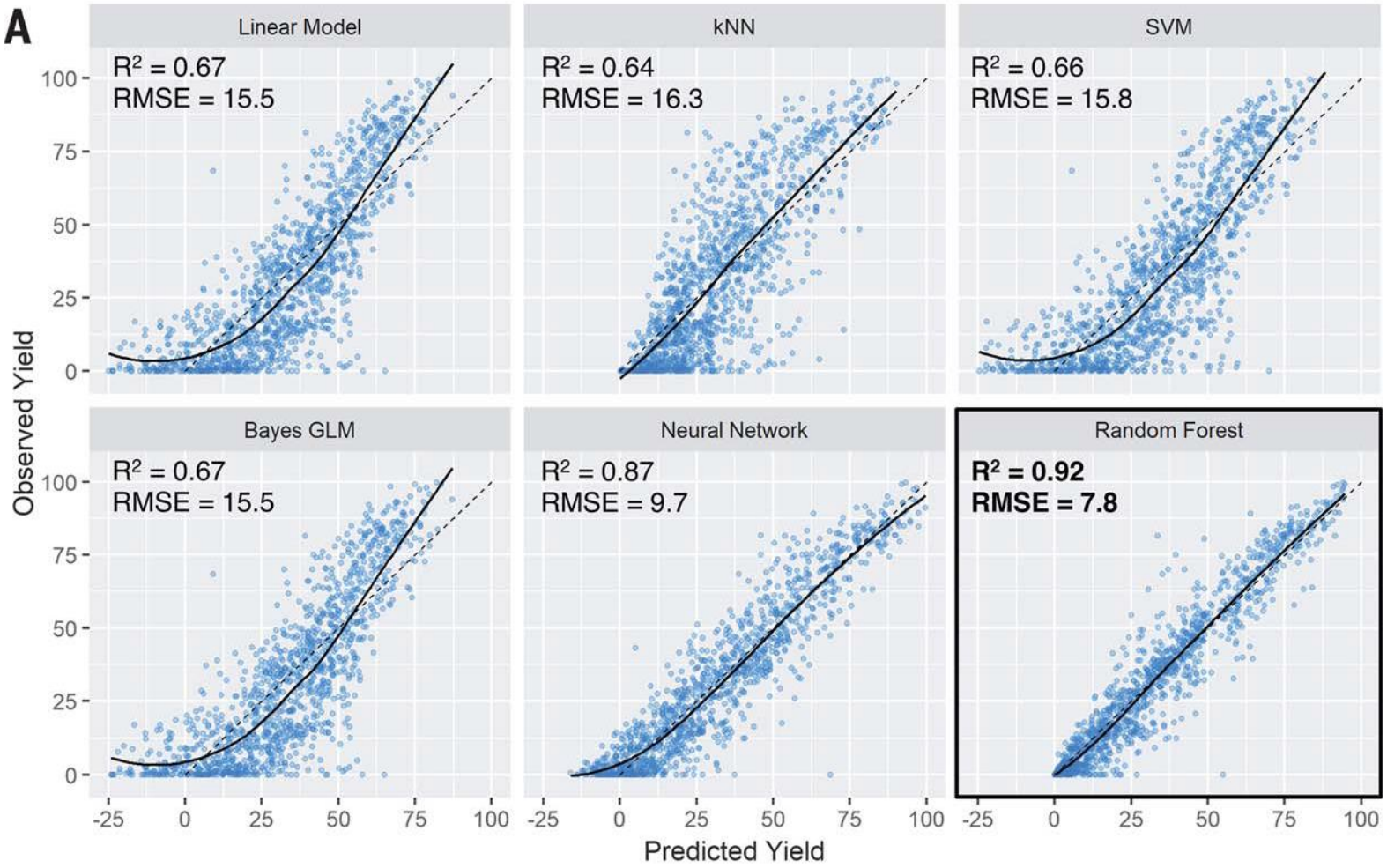
30% → test set
Prediction with calculated model was evalueted.
(R^2, root mean square error(RMSE))

Training set and test set must be differrent to make reliable algorithm.



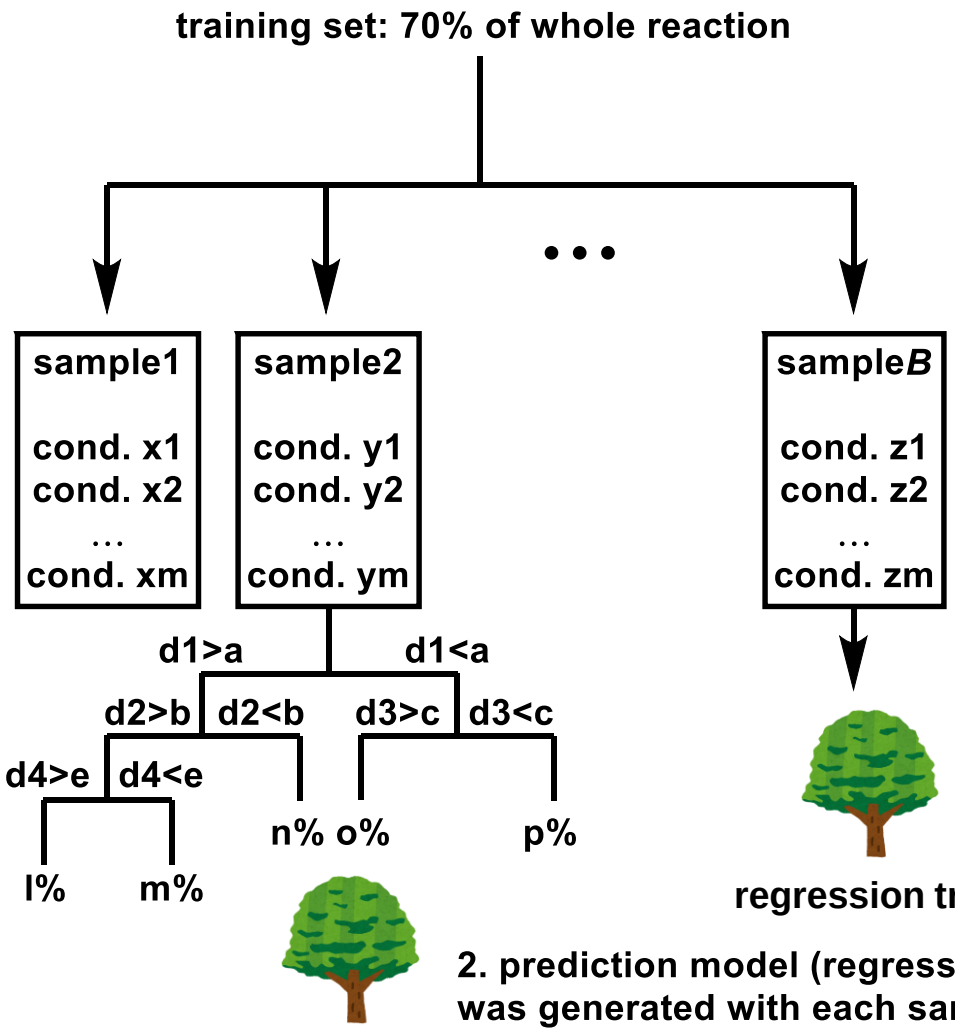
1. Split into 70:30
2. Learn from 70%
3. Test with 30%
4. Evaluate

Prediction Results



Random forest model was best fitting.

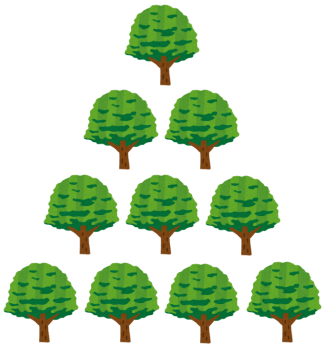
Random Forest Model



1. make *B* (bootstrap)samples (duplication allowed)

2. prediction model (regression tree) was generated with each sample

3. results from randomly generated trees was combined (average value is the predict yield)
high predictive algorithm obtained bycombining many low-prcision models



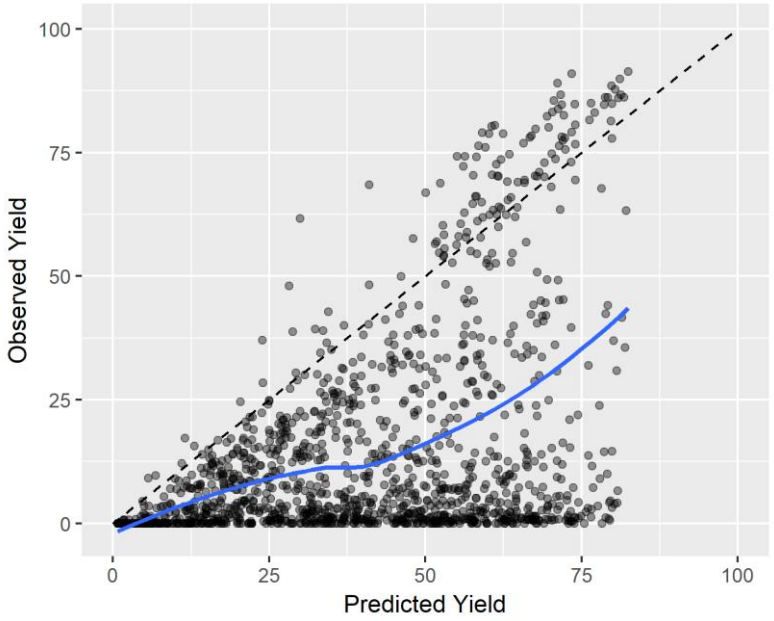
random forest

d1, d2, d3, d4: randomly chose descriptors

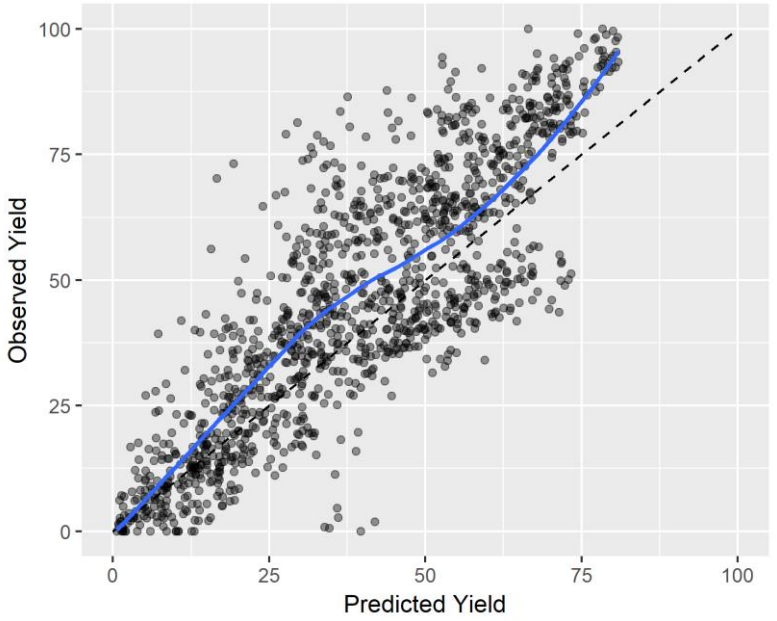
See also 160723_LS_Shunichiroh_KATOH

Problems in ML

“Activity cliffs”
…modest changes in chemical structure can lead to notable changes in reaction outcome
(Cruz-Monteagudo, M. *et al.*, *Drug Discov. Today* 2014, *19*, 1069.)



Training: aryl bromide
Test: aryl chloride
-> overpredicted

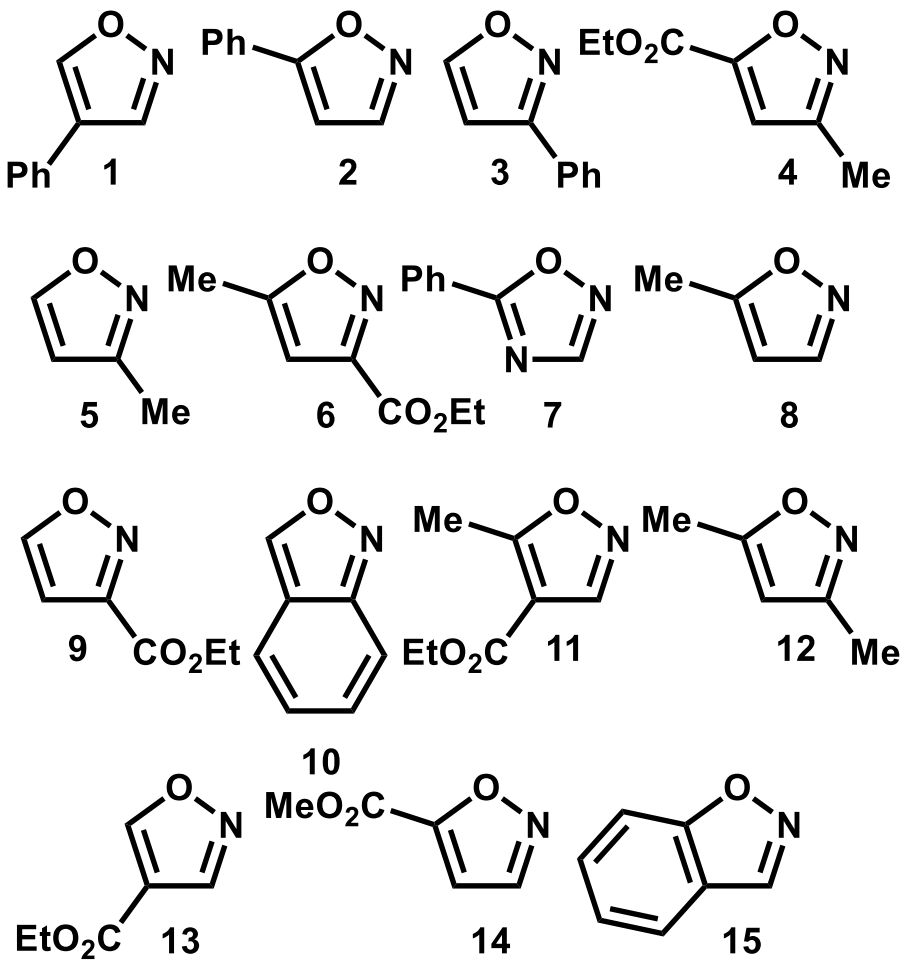


Training: aryl bromide
Test: aryl iodide
-> underpredicted

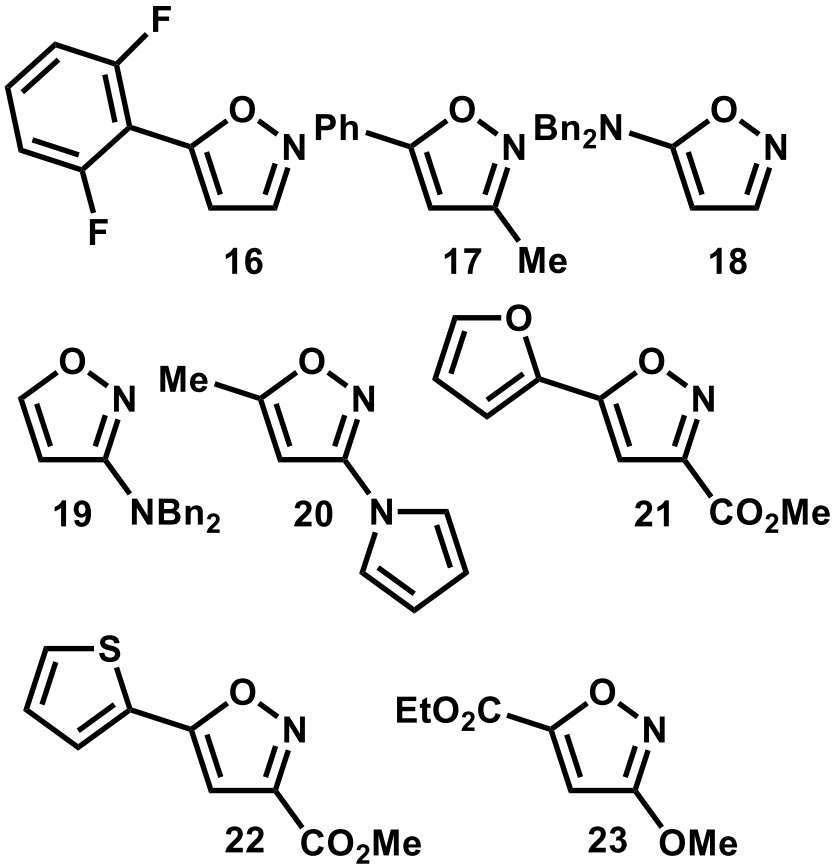
→training data need to spread across the chemical space of interest.

Out-of-Sample Prediction

Training set

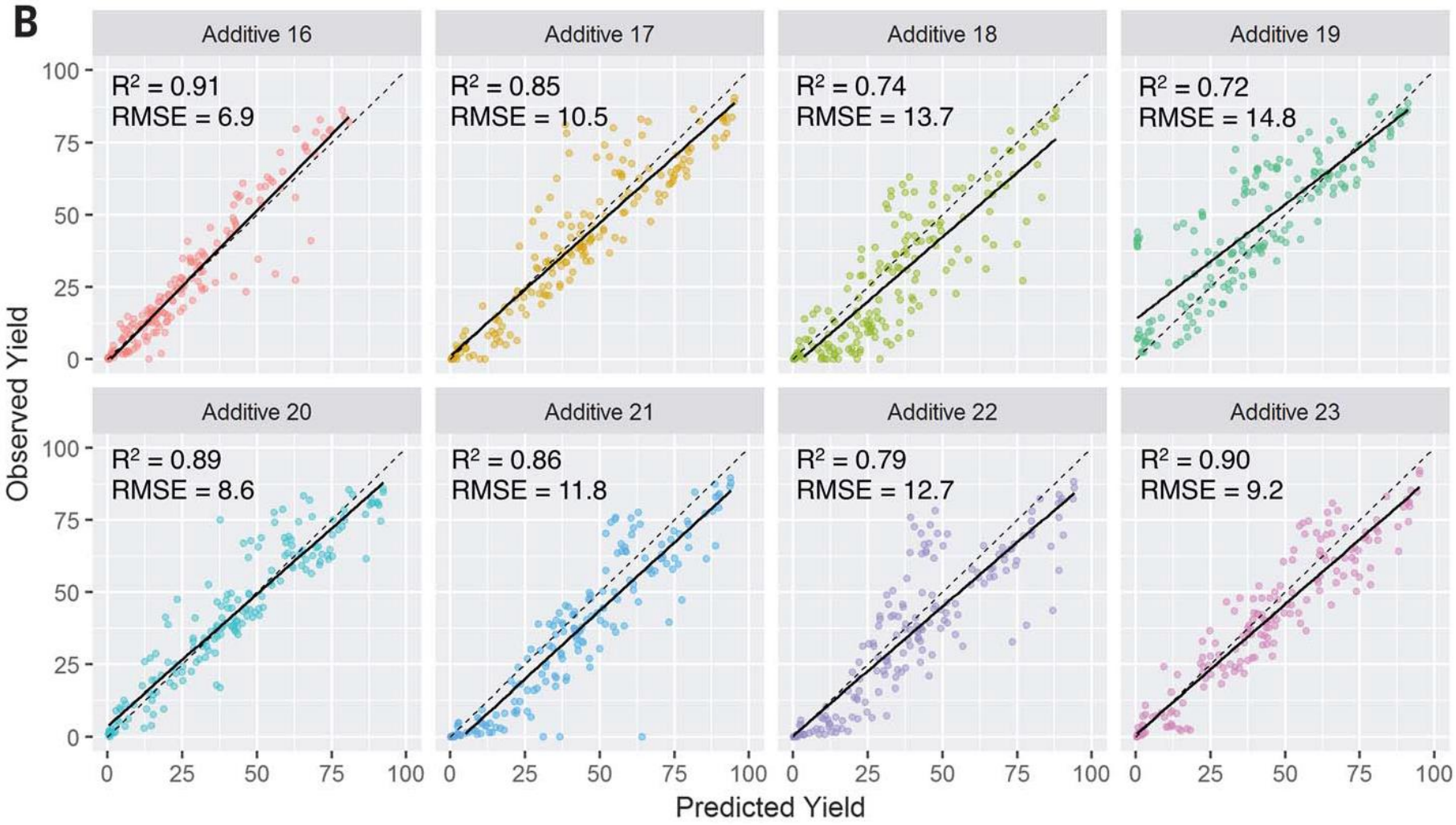


Test set



Out-of-Sample Prediction

Train: 15 additives → Test: other 8 additives



Mechanistic Analysis

It is difficult to interpret the outcome of ML.
Which is the important descriptor???

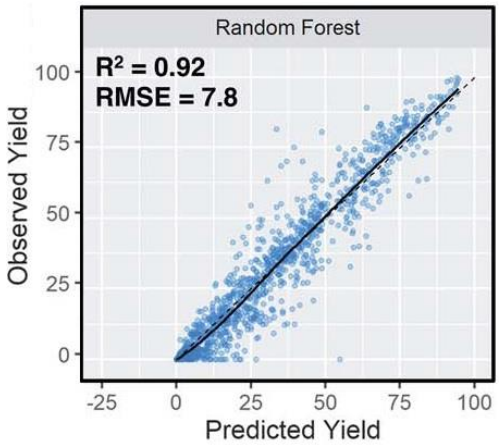
Conditions	Descriptor1	Descriptor2	…
1	X_1	Y_1	…
2	X_2	Y_2	…
3	X_3	Y_3	…
…	…	…	…

→

Conditions	Descriptor1	Descriptor2	…
1	X_l	Y_1	…
2	X_m	Y_2	…
3	X_n	Y_3	…
…	…	…	…

Descriptor1 randomized

↓ **prediction**



→ **Prediction get worse**

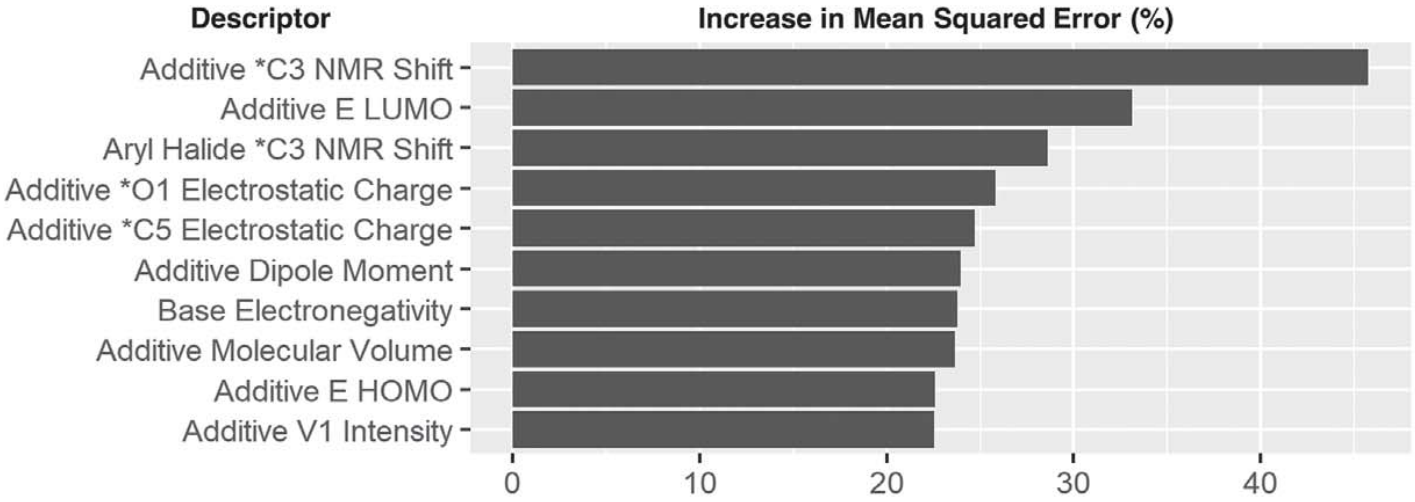
↓ **prediction**

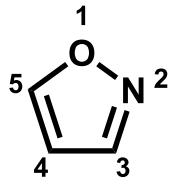
How much the mean squared error(MSE) increased.

If MSE increases a lot
… the randomized descriptor is important.

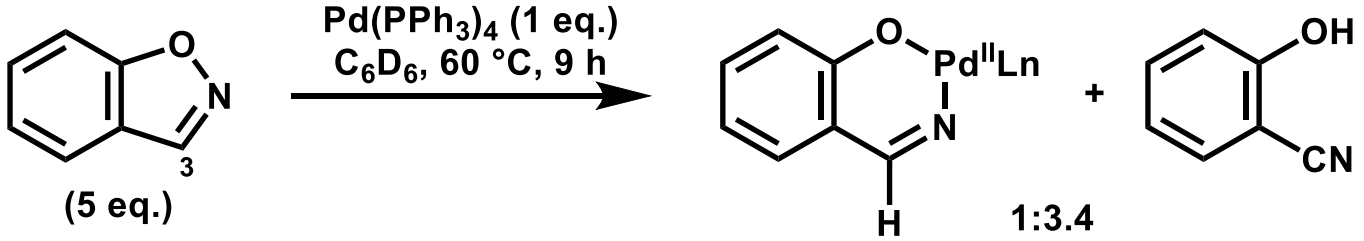
1) T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction” *Springer*, **2009**.

Variable Importance Plot





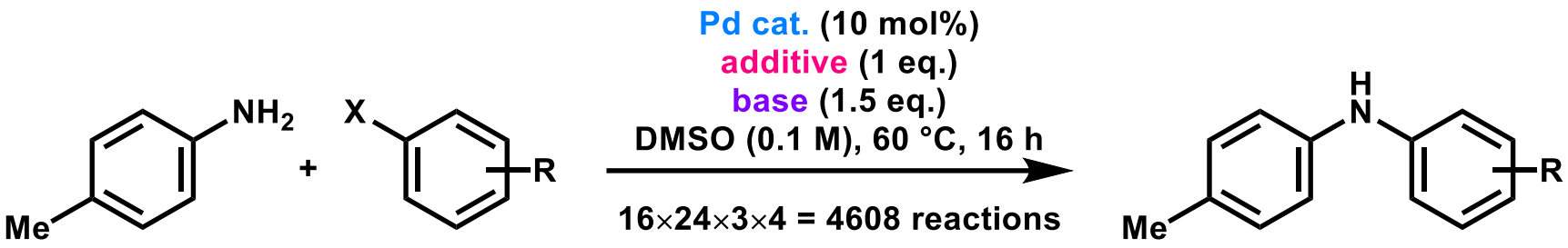
Authors said that "the descriptors suggest that the propensity of the additive to act as an electrophile influences reaction outcomes."
So author hypothesized oxidative addition of Pd to isoxazole was main side reaction, and it was confirmed by experiment.

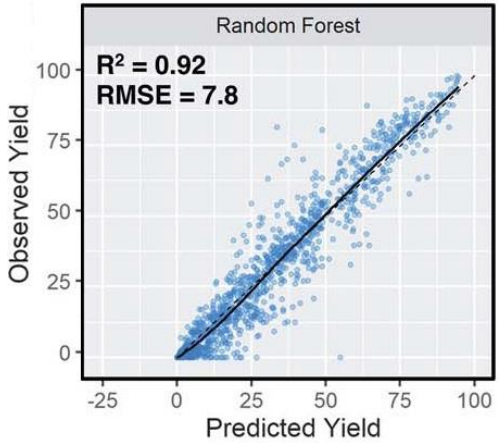


However, it is difficult to connect *C3 shift and oxidative addition.

Problem in this method is that if multiple correlated variables are present, each will individually appear to have less importance.

Short Summary





The result of intermolecular reaction screening was predicted.

However, as additive consumption was not included in out put, the algorithm will have problem when substrates with embedded isoxazoles.

✔ **Yield was predicted with high accuracy.**
✔ **All descriptor was calculated only by computer.**
✔ **Reaction performance can be predicted without chemical theory.**

- **Interpretation of valuable importance plot.**
- **In this report, aryl halides and isoxazoles were used as screening conditions. These compound has large similarity in chemical structures.**

Contents

1. Introduction
2. Prediction in Intermolecular Reaction Screening (main paper)
3. **Prediction in Usual Reaction Screening**

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Communication

Cite This: *J. Am. Chem. Soc.* XXXX, XXX, XXX–XXX

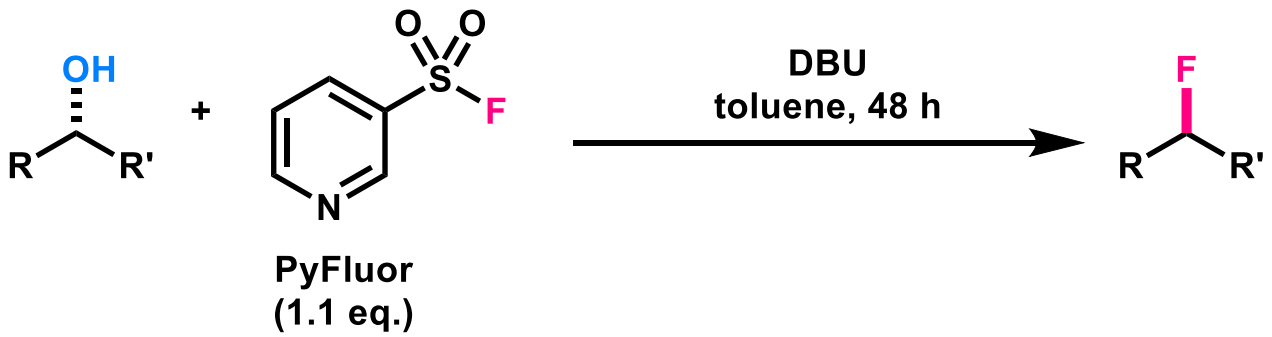
pubs.acs.org/JACS

Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning

Matthew K. Nielsen, Derek T. Ahneman, Orestes Riera, and Abigail G. Doyle*

Dehydroxyfluorination

Nielsen, M. L.; Ugaz, C. R.; Li, W.; Doyle, A. G. *J. Am. Chem. Soc.* **2015**, 137, 9571



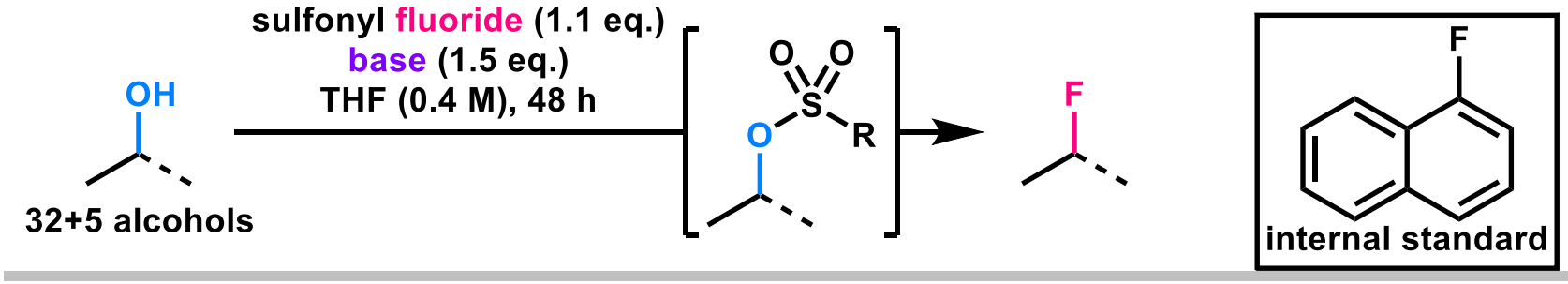
PyFluor
(1.1 eq.)

thermally stable, inexpensive
diversity in sulfonylfluorides

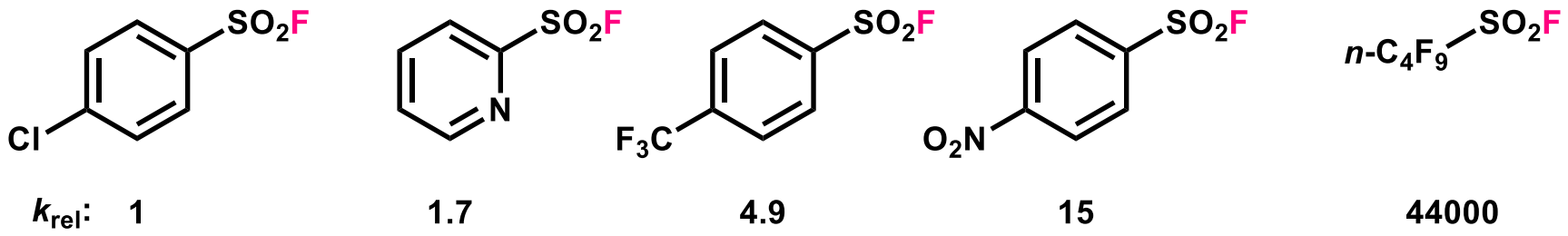
The system is useful for acyclic secondary alcohol.
This protocol gave poor yield for activated alcohols, because the intermediate were attacked by DBU.
If the reaction didn't proceed well with "best condition", alternative condition must be investigated.

-> Reaction screening was conducted to understand the reactivity of sulfonyl fluorides.

Reaction Screening



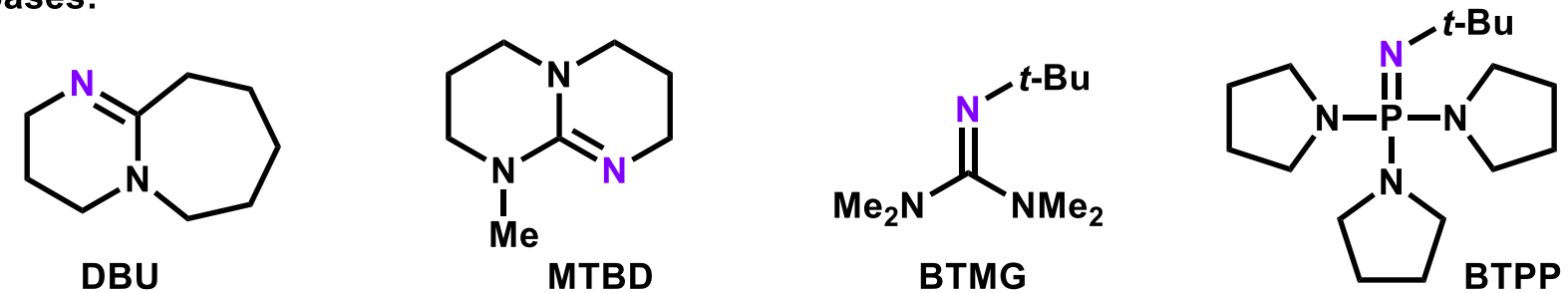
5 sulfonyl fluorides:



k_{rel}: 1 | 1.7 | 4.9 | 15 | 44000

reactive

4 bases:

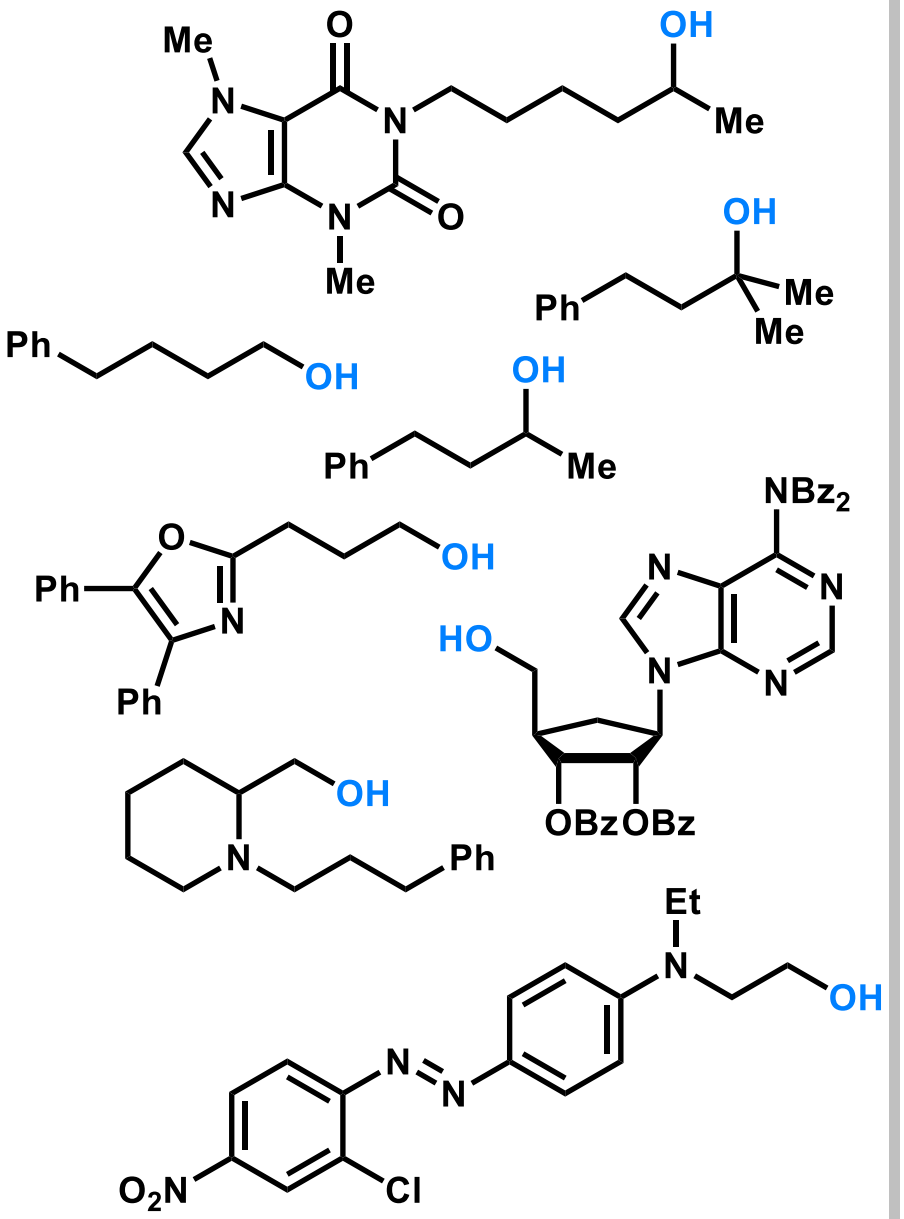


sterically bulky

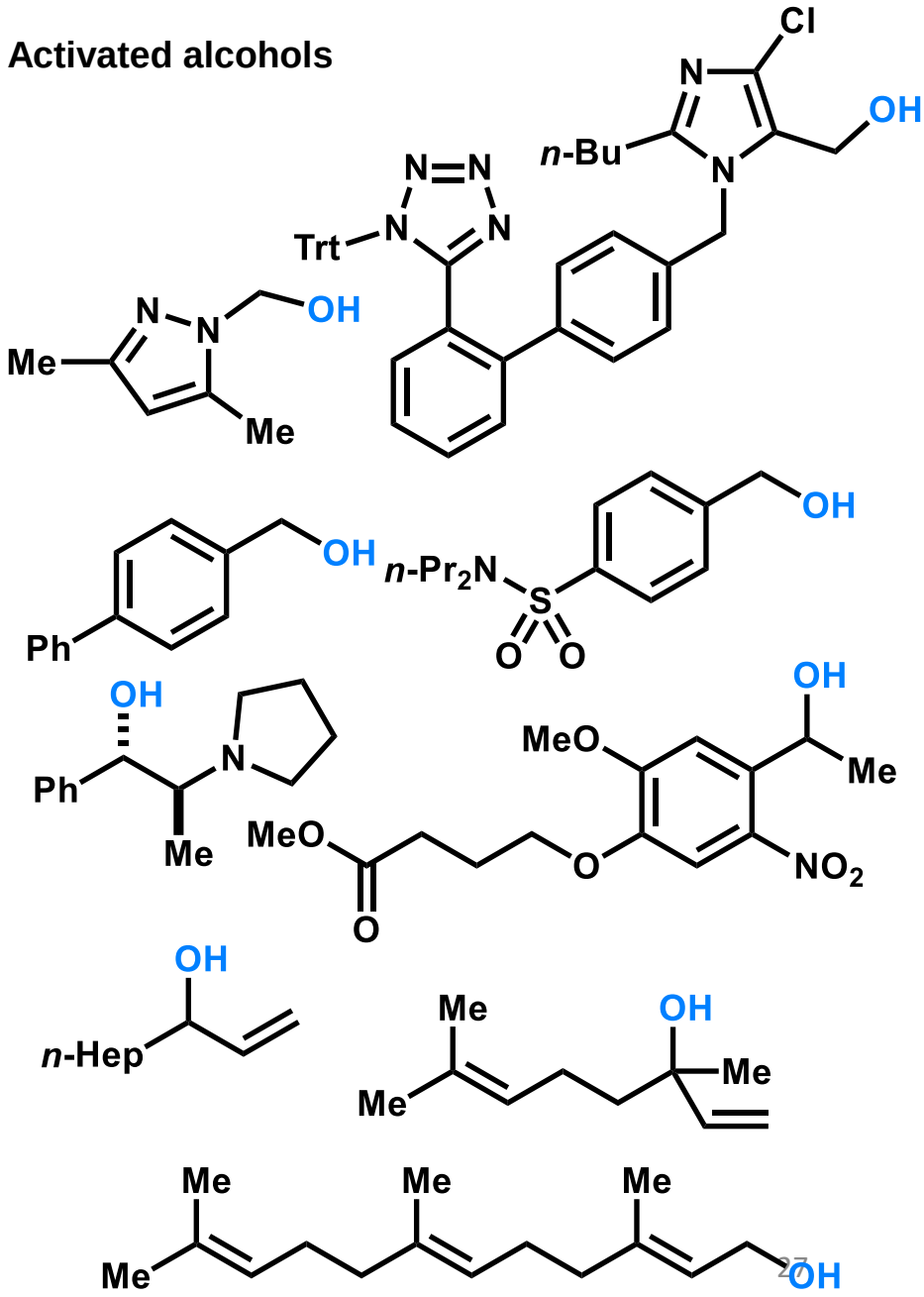
32×5×4 = 640 reactions

Alcohols

Unctivated alcohols

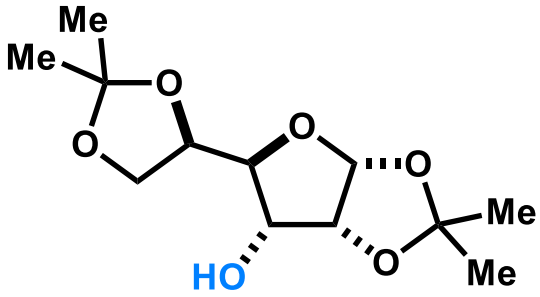


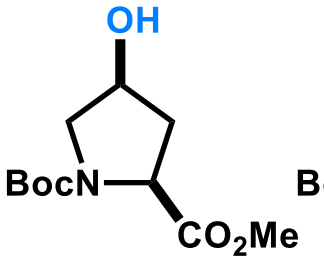
Activated alcohols

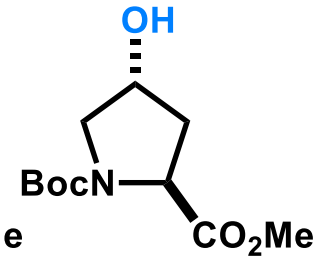


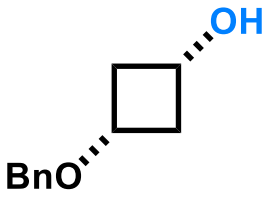
Alcohols

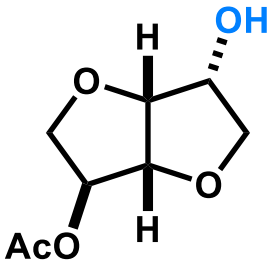
Cyclic alcohols

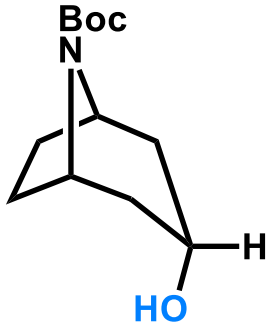


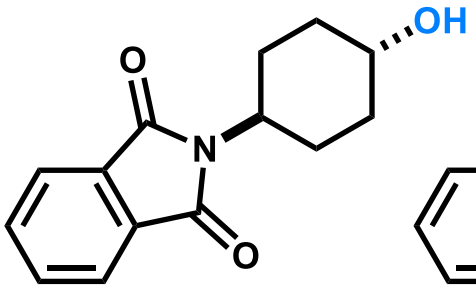


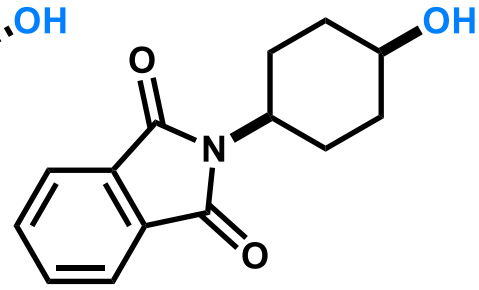




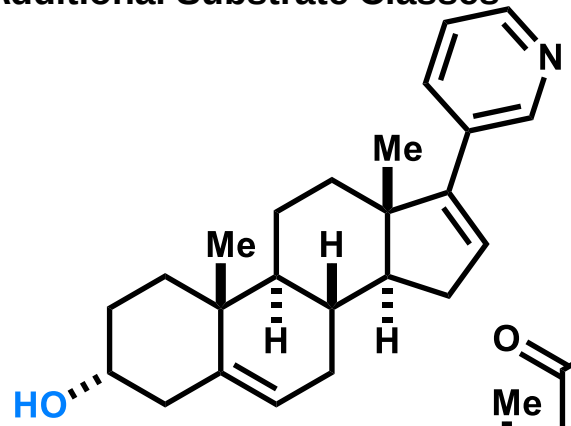


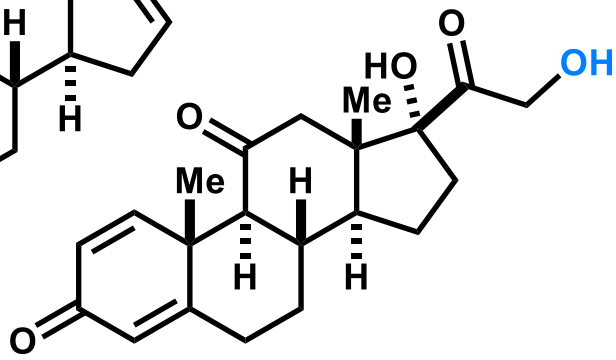


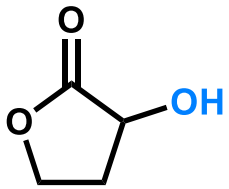


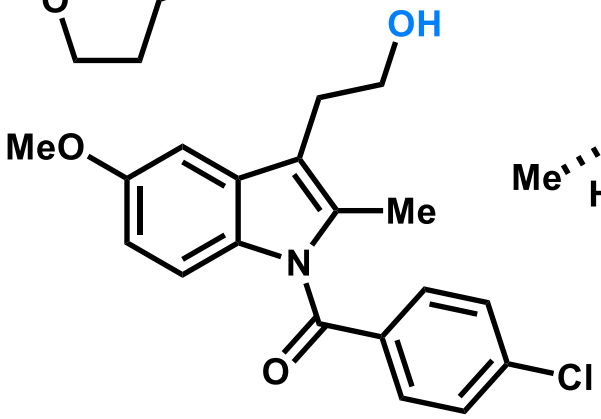


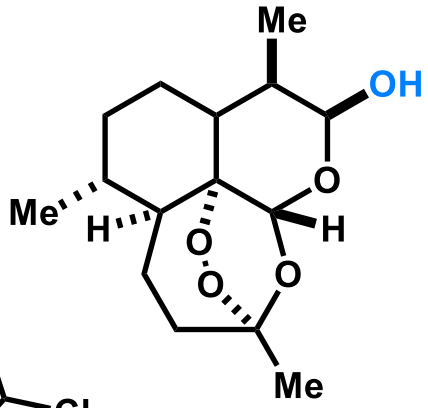
Additional Substrate Classes

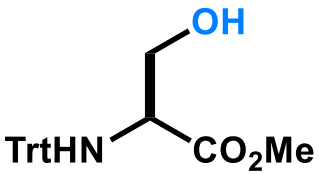


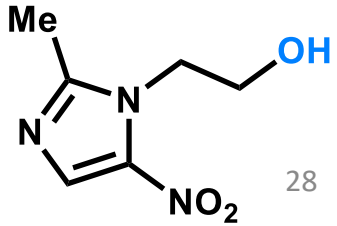




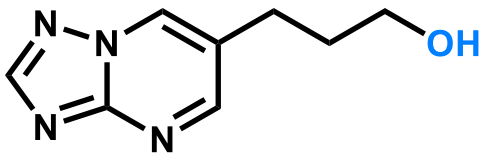




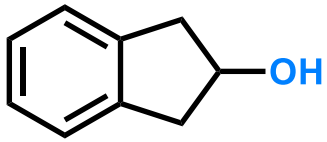




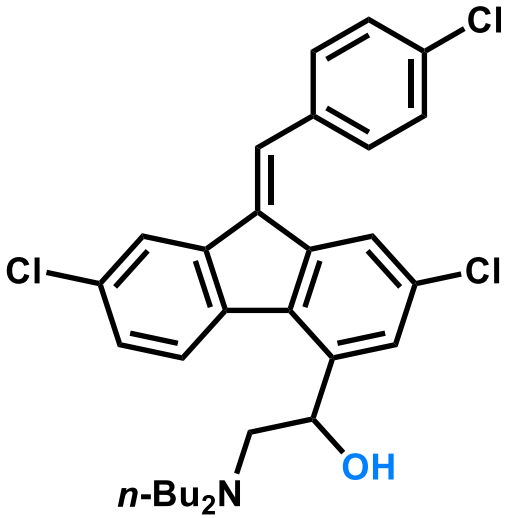
External Validation Set



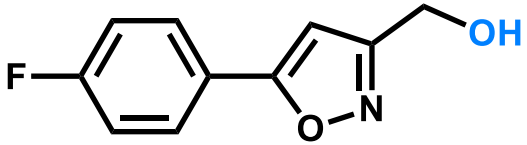
1



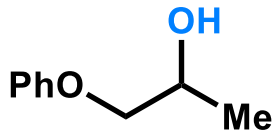
2



5

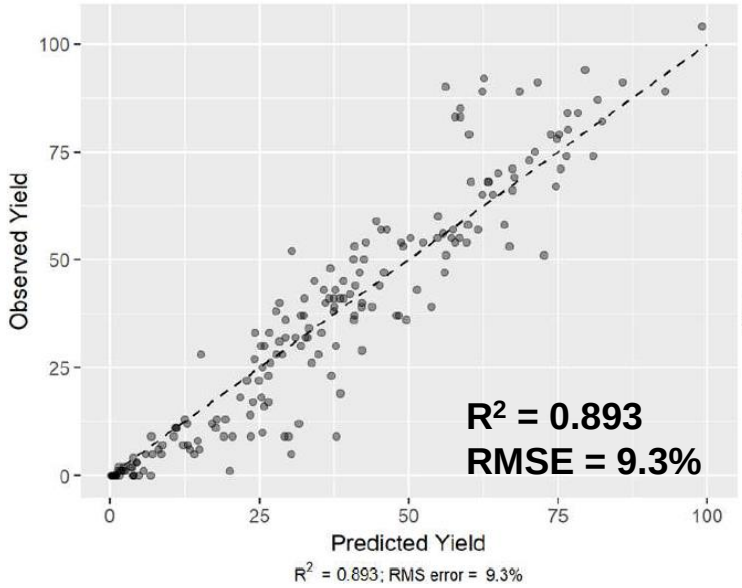


3

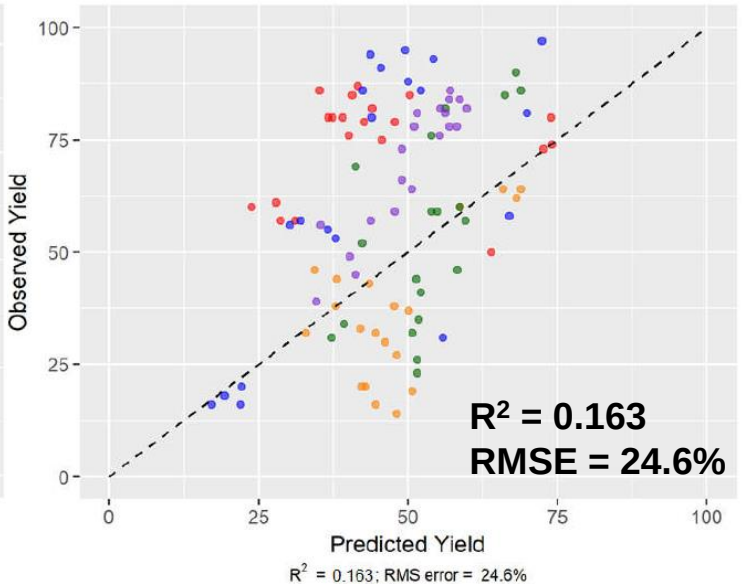


4

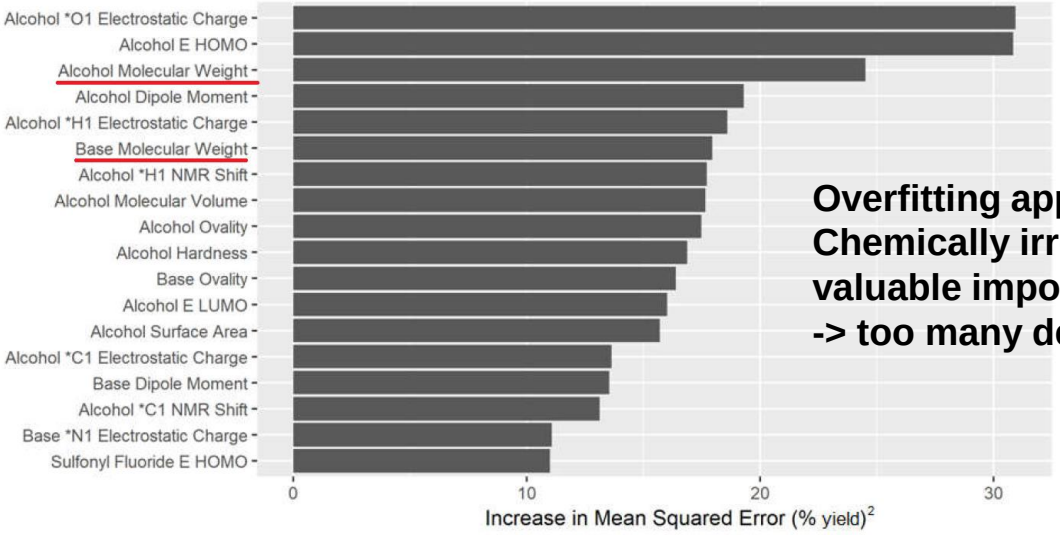
Problem in Previous Method



Prediction of 30% test set



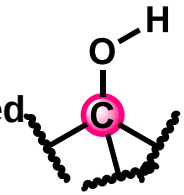
Prediction of external validation set
-> poor prediction



Overfitting appeared
Chemically irrelevant variables are highlighted in valuable importance plot (left figure).
-> too many descriptor makes the performance worse.

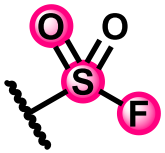
Modified Method

Descriptors relevant to the nucleophilic substitution was selected and new categorical descriptors was added.



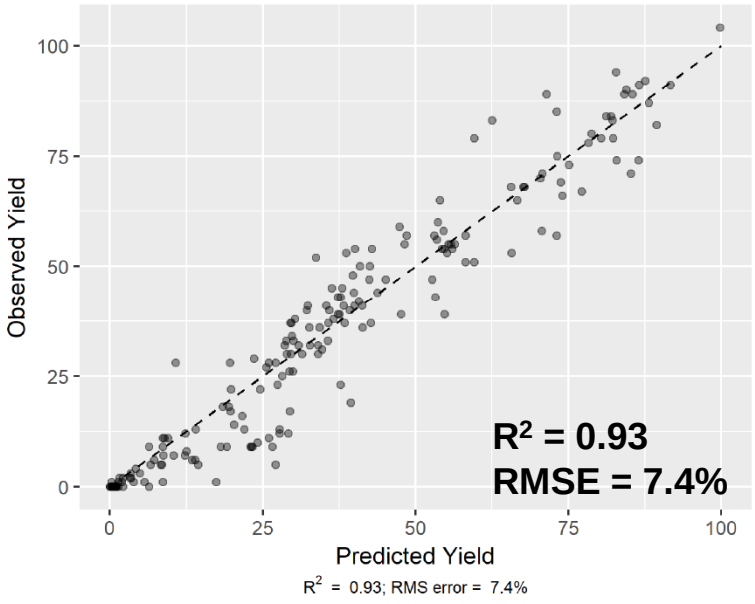
alcohol

base

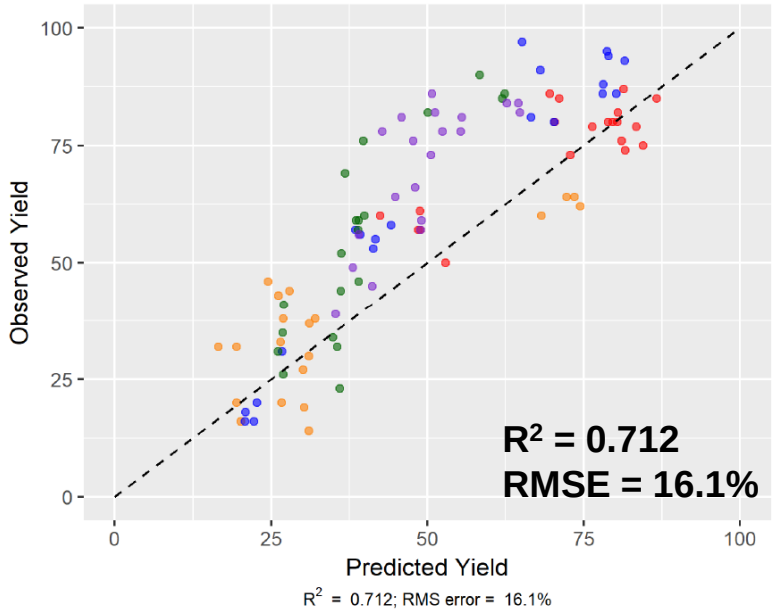


sulfonyl fluoride

- **Calculated descriptors**
  - **Alcohol - *C1 electrostatic charge/*C1 exposed area/electronegativity**
  - **Base - *N1 exposed area**
  - **Sulfonyl fluoride - *S1 electrostatic charge/*F1 electrostatic charge/*O1 electrostatic charge**
- **Categorical descriptors**
  - **Alcohol – primary/secondary/tertiary/cyclic/4-membered ring/5-membered ring/6-membered ring/7-membered ring/benzylic/allylic/homobenzylic/homoallylic/α-carbonyl/β-carbonyl/hemiacetal/amino alcohol**



Prediction of 30% test set

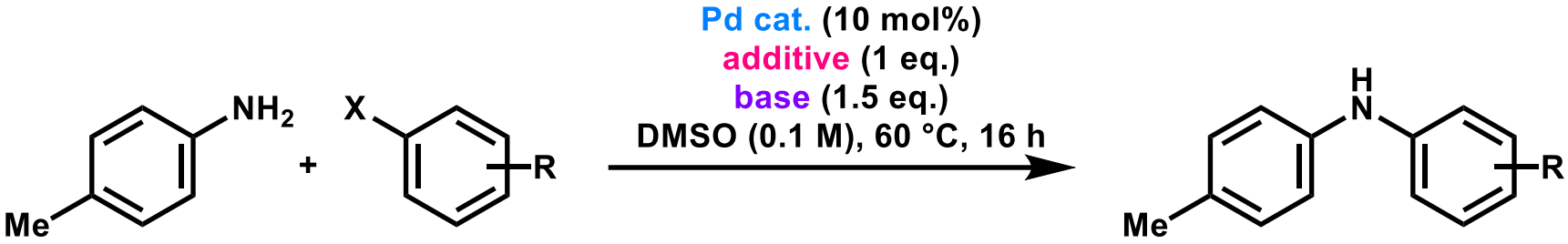


Prediction of external validation set

Summary

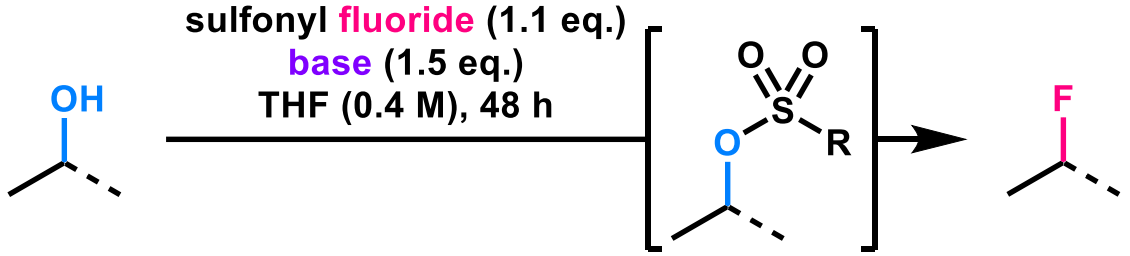
Machine Learning was applyed to predict reaction performance

1. Intermolecular reaction screening



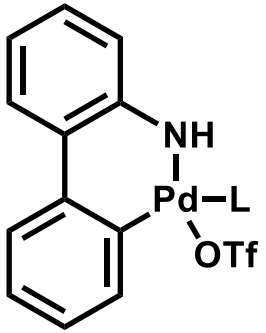
Reaction perfomance can be predicted well by automated caluculation.

2. Usual substrate scope

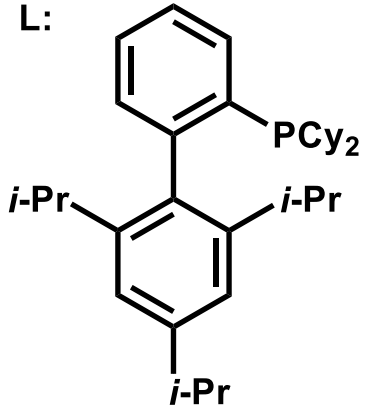


Reaction perfomance can be predicted well by defining appropriate descriptor.

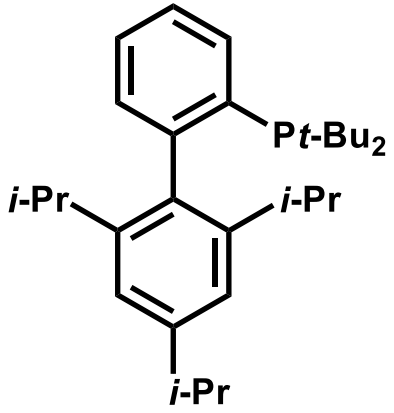
Ligands



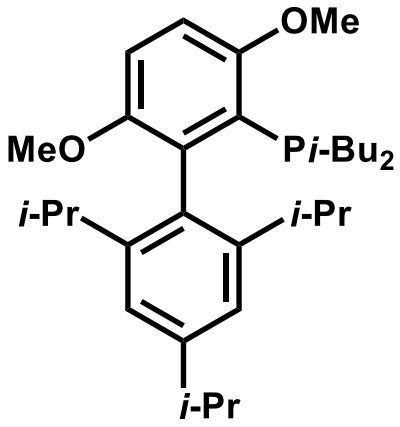
L:



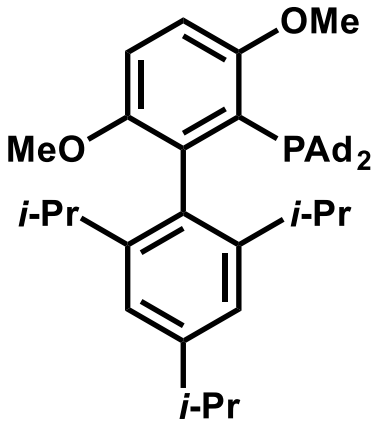
XPhos



t-BuXPhos



t-BuBrettPhos



AdBrettPhos

Descriptors

- **Additive Descriptors (*n* = 19)**

E_{HOMO}, E_{LUMO}, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular Weight, Ovality, Surface Area, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, *C4 Electrostatic Charge, *C5 NMR Shift, *C5 Electrostatic Charge, *N1 Electrostatic Charge, *O1 Electrostatic Charge, V1 Frequency, V1 Intensity

- **Aryl Halide Descriptors (*n* = 27)**

E_{HOMO}, E_{LUMO}, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular Weight, Ovality, Surface Area, *C1 NMR Shift, *C1 Electrostatic Charge, *C2 NMR Shift, *C2 Electrostatic Charge, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, *C4 Electrostatic Charge, *H2 NMR Shift, *H2 Electrostatic Charge, *H3 NMR Shift, *H3 Electrostatic Charge, V1 Frequency, V1 Intensity, V2 Frequency, V2 Intensity, V3 Frequency, V3 Intensity

- **Base Descriptors (*n* = 10)**

E_{HOMO}, E_{LUMO}, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular Weight, Ovality, Surface Area, *N1 Electrostatic Charge

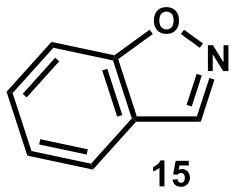
- **Ligand Descriptors (*n* = 64)**

Dipole Moment, *C1 NMR Shift, *C1 Electrostatic Charge, *C2 NMR Shift, *C2 Electrostatic Charge, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, *C4 Electrostatic Charge, *C5 NMR Shift, *C5 Electrostatic Charge, *C6 NMR Shift, *C6 Electrostatic Charge, *C7 NMR Shift, *C7 Electrostatic Charge, *C8 NMR Shift, *C8 Electrostatic Charge, *C9 NMR Shift, *C9 Electrostatic Charge, *C10 NMR Shift, *C10 Electrostatic Charge, *C11 NMR Shift, *C11 Electrostatic Charge, *C12 NMR Shift, *C12 Electrostatic Charge, *C13 NMR Shift, *C13 Electrostatic Charge, *C14 NMR Shift, *C14 Electrostatic Charge, *C15 NMR Shift, *C15 Electrostatic Charge, *C16 NMR Shift, *C16 Electrostatic Charge, *C17 NMR Shift, *C17 Electrostatic Charge, *H11 NMR Shift, *H11 Electrostatic Charge, *H3 NMR Shift, *H3 Electrostatic Charge, *H4 NMR Shift, *H4 Electrostatic Charge, *H9 NMR Shift, *H9 Electrostatic Charge, *P1 Electrostatic Charge, V1 Frequency, V1 Intensity, V2 Frequency, V2 Intensity, V3 Frequency, V3 Intensity, V4 Frequency, V4 Intensity, V5 Frequency, V5 Intensity, V6 Frequency, V6 Intensity, V7 Frequency, V7 Intensity, V8 Frequency, V8 Intensity, V9 Frequency, V9 Intensity, V10 Frequency, V10 Intensity

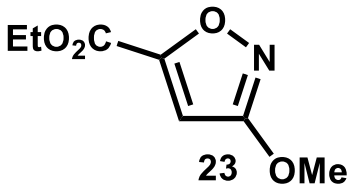
Isoxazole Descriptors

No.	*C3_NMR_shift	*C3_electrostatic_charge	*C4_NMR_shift	*C4_electrostatic_charge	*C5_NMR_shift	*C5_electrostatic_charge	*N1_electrostatic_charge	*O1_electrostatic_charge	E_HOMO	E_LUMO
15	139.05	0.009	116.54	0.009	154.93	0.429	-0.25	-0.158	-0.2485	-0.0459
23	163.09	0.721	95.32	-0.581	155.72	0.241	-0.383	-0.106	-0.2605	-0.0587

No.	V1_frequency	V1_intensity	dipole_moment	Electronegativity	hardness	molecular_volume	molecular_weight	Ovality	surface_area
15	892.395	10.204	3.184221	0.15	0.1	121.96	119.123	1.152	137.07
23	697.362	5.037	3.44567	0.16	0.1	164.92	171.152	1.343	195.35



15

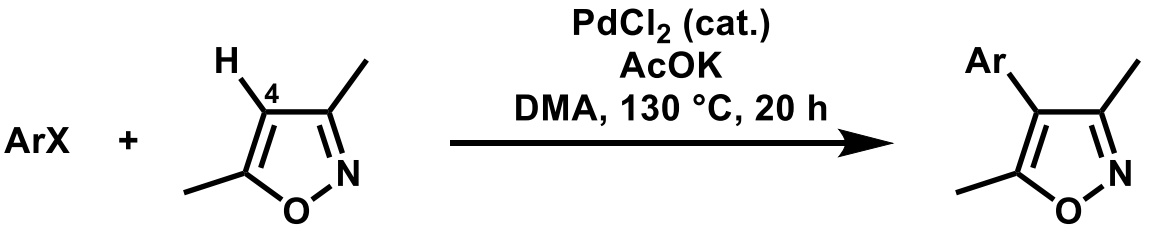


23

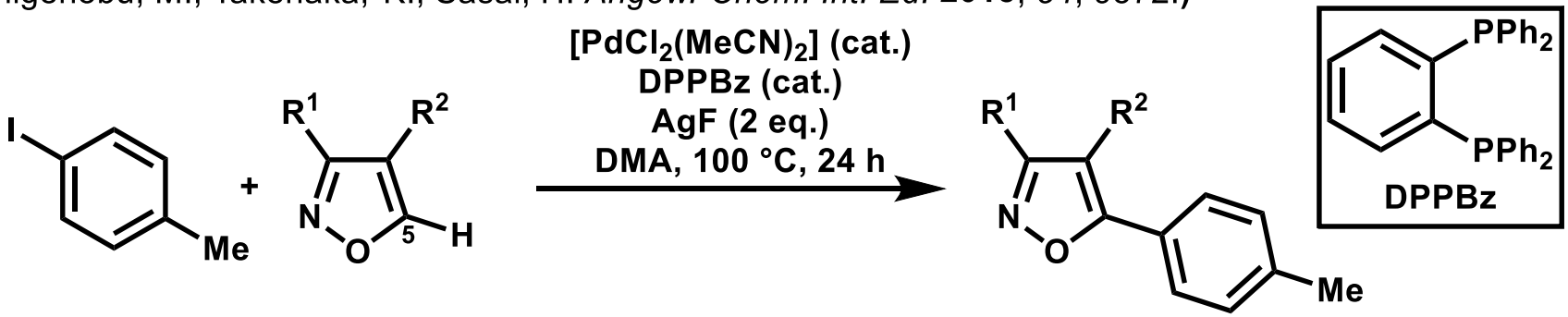
Reaction of Isoxazole and Transition Metal

• **Direct C-H arylation of Isoxazoles**

(Fall, Y.; Reynaud, C.; Doucet, H.; Santelli, M. *Eur. J. Org. Chem*. **2009**, 4041.**)**

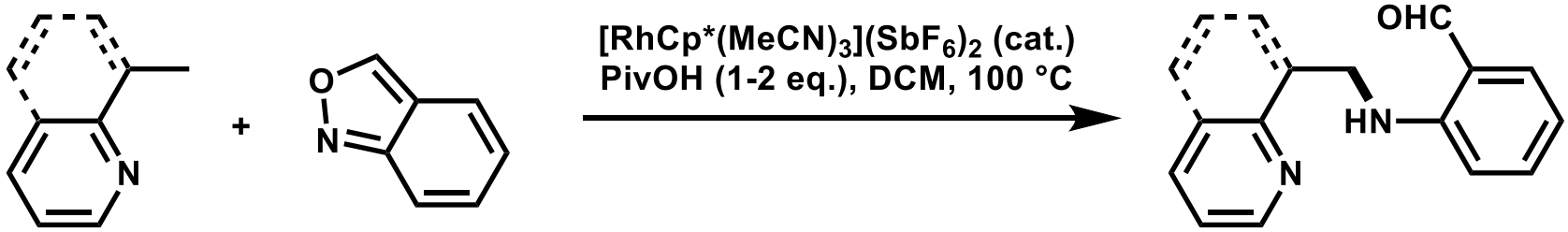


(Shigenobu, M.; Takenaka, K.; Sasai, H. *Angew. Chem. Int. Ed.* **2015**, *54*, 9572.**)**

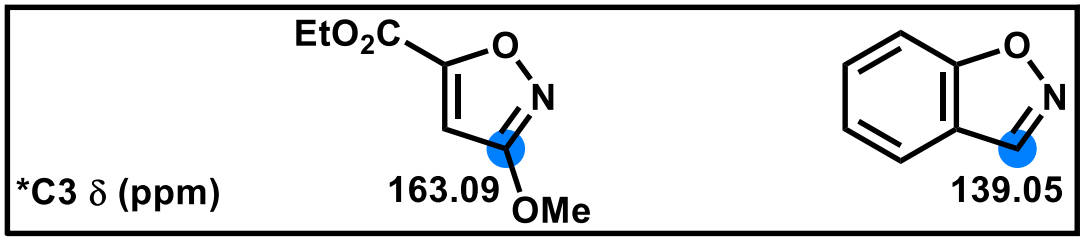


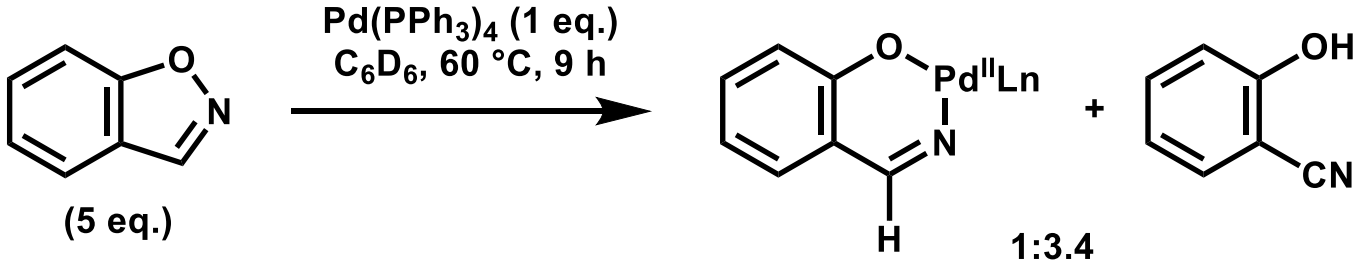
• **Oxidative addition of trinsition metal to isoxazole N-O bond**

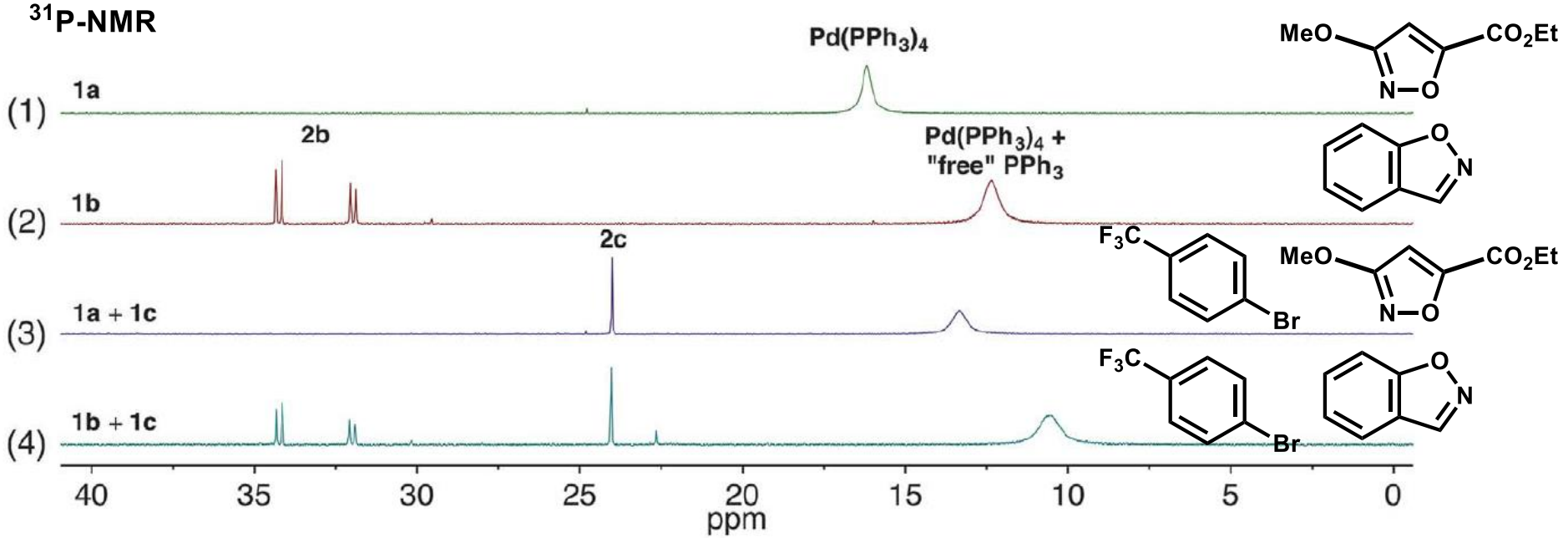
(Yu, S.; Tang, G.; Li, Y.; Zhou, X.; Lan, Yu.; Li, X., *Angew. Chem. Int. Ed.* **2016**, *55*, 8696.**)**



Reaction of isoxazole with $Pd(PPh_3)_4$







Side reaction was predicted and determined by ML.
Important descriptor was extracted automatically by ML.

Combined Graph (Fluorination)