Baran Lab
Group Meeting

Statistics in Synthetic Chemistry

Marcus Farmer
09/02/2017

Central goal of this group meeting

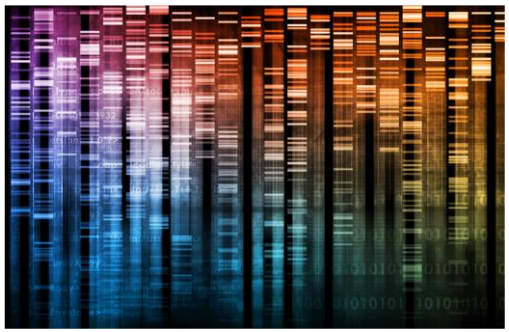
"To provide a *brief* summary of statistical methodology that has found applications in organic chemistry, as well as to provide a forward looking perspective of potential applications that have yet to be realized. Emphasis will be placed on how to utilize statistical methodology so that the aforementioned potential applications may be readily sought after by organic chemists."

Select applications of statistics in areas related to synthesis

Clinical Trials

How sure can I be that 900 people is a sufficient sample size to study the impact of this drug on a disease?

How do I deterimine what dose to conduct this new clinical trial at?



Biostatistics/Genomics

How do I automatically classify genetic variations?

Is there a gene that influences a person's response to a drug that we didn't realize before?

Select applications of statstics in areas not related to synthesis

LAS VEGAS

What is statistics?

"Statistics is a science in my opinion, and it is no more a branch of mathematics than are physics, chemistry, and economics; for if its methods fail the test of experience - not the test of logic- they are discarded"
- John Tukey
Founding Chairman of Princeton's Statistics Department

What are statistical parameters?

Statistical parameters are theoretical values that we can only approximate by using statistics.

What's the difference between probability and proportion?

Probability is a statistical parameter that can be estimated by the ratio of occurences, i.e. the proportion.

What are confidence intervals and what do they tell us?

Confidence intervals are used to declare ranges where we can be certain, with some percentage of confidence, the true value of the statistical parameter exists.

What is statistical power?

Statistical power is the confidence we have that our test will be able to distinguish the difference between two statistics

What is correlation and how should we use it properly?

Correlation should not be mistaken for causation and is merely the description of an apparent relatonship between variables.

What is Bayes theorem and why should I care?

Bayes theorem details how we should update our beliefs about the probability of an event given new information on the system.

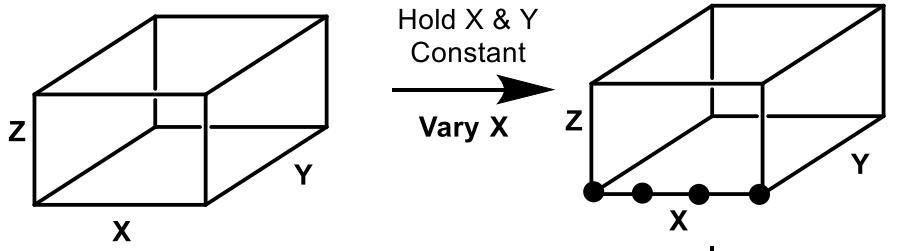
Where can I learn more about statistics?

Penn State's online world campus is a fantastic place to start
Link: **https://onlinecourses.science.psu.edu/statprogram/**

Design of Experiments (DoE): Principles and Theory

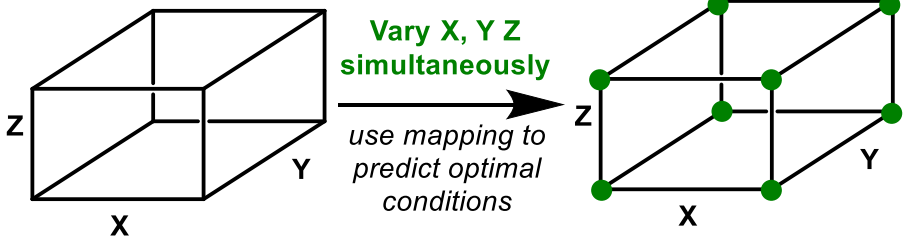
Types of Experiments

One Variable at a Time (OVAT)

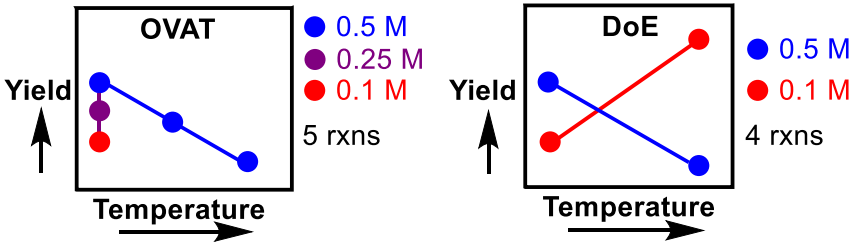


This is the traditional optimization scheme used by organic chemists. This example has 10 unique reactions, but are we sure that we found the optimal conditions? Note that yield is the 4th dimension in this experiment.

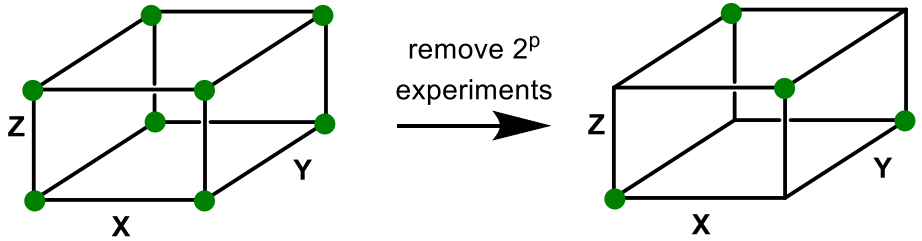
Factorial Design (DoE)



DoE vs OVAT displayed in 2-D (function space is 3D)



Fractional Factorial Design



Factorial design (2^3) — Fractional Factorial Design (2^{3-1})

General form of a factorial experiment: L^{F-P}

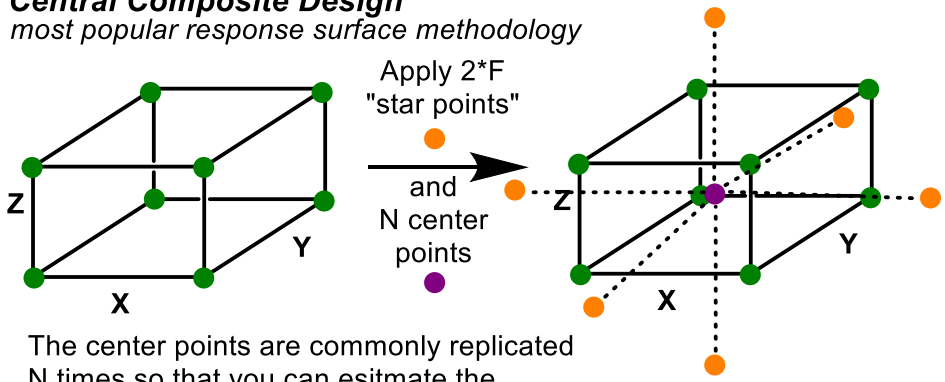
Where L = levels, F = factors, and P = number used to reduce complexity

Response Surface Methodology

Factorial designs are great for quick experiments and will tell you if there are any interactions. However, it won't necessarily let you model the shape of the "response surface." Response surface methodology tackles this by teaching you how to design experiments that allow you to model the "response surface" and approximate it's curvature.

Central Composite Design

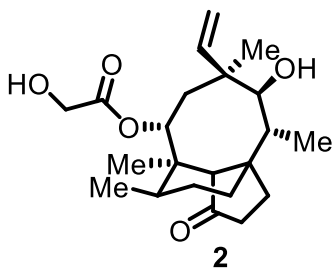
most popular response surface methodology



The center points are commonly replicated N times so that you can esitmate the variation of your response surface

What will you need? *Any of the following software will be helpful.*
DoE Software: Design Expert (Stat-Ease Inc.), MODDE (Umetrics), DoE Fusion PRO (S-Matrix Corp.), STAVEX (Aicos), Minitab (Minitab Inc.), JMP (SAS).
Free DoE Software: R (packages can be found in the CRAN. Requires programming skills.)
Where to learn more? *Penn State Online STA 503 (Free Notes!)*

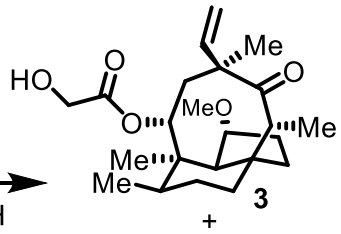
Design of Experiments (DoE): A Case Study of a GSK Process Optimization



2

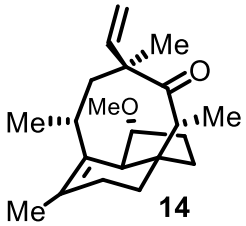
H_2SO_4

TMOF, MeOH



3

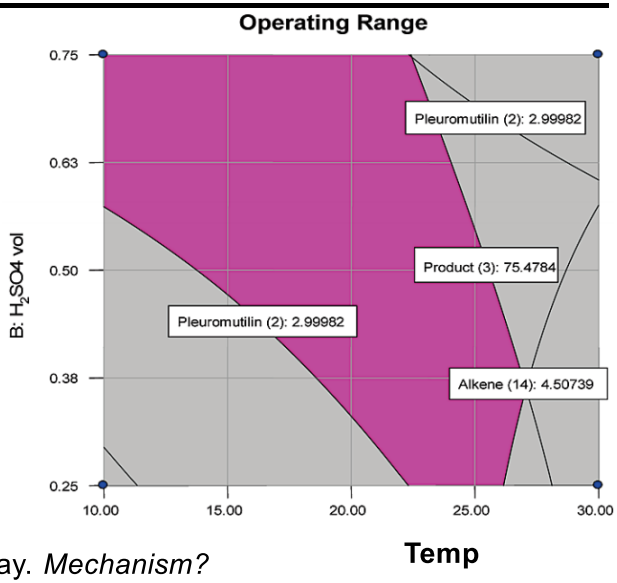
+



14

Study Objectives
1) Maximize **3**
2) Minimize **2**
3) Minimize **14**

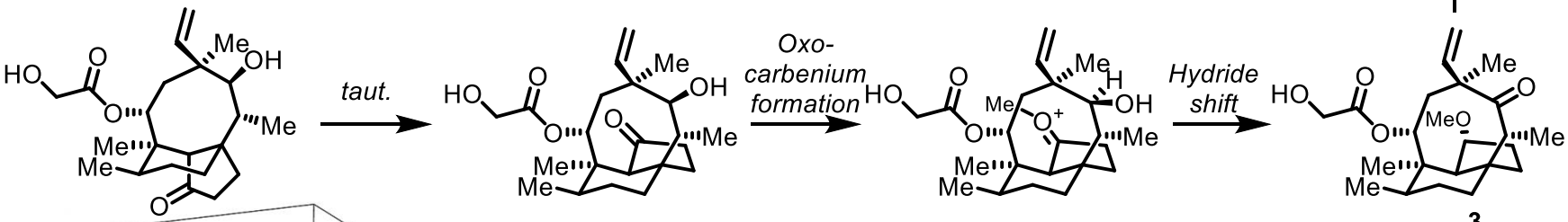
Experimental Design
1) Half Factorial
2) RSM
(central composite,
30 experiments)

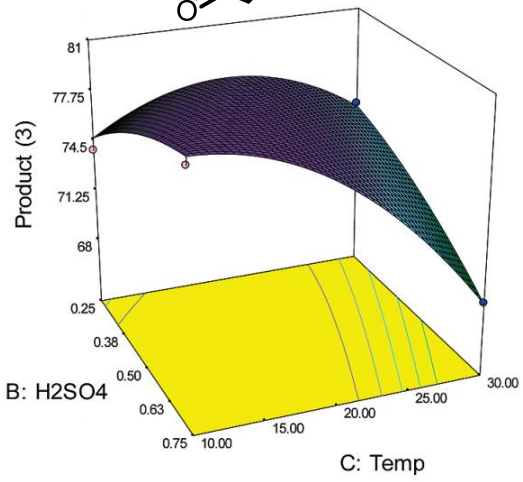


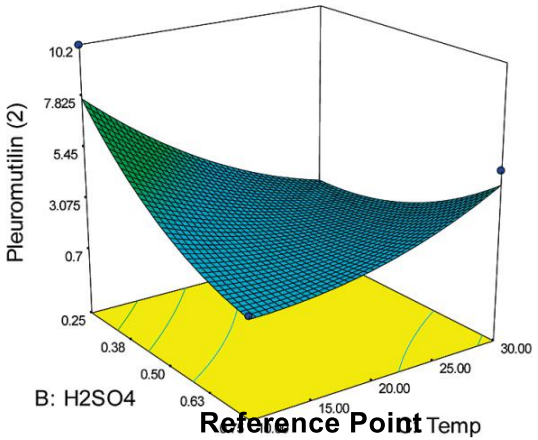
GSK *Org. Proc. Res. Dev.* **2009**, *13*, 729.

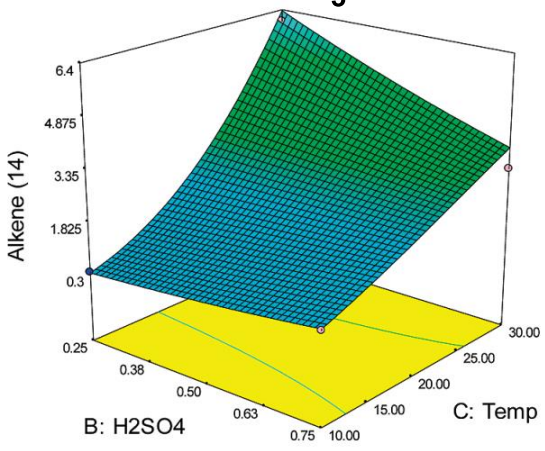
factor	initial set-points	ranges (half-factorial study)	ranges (RSM study)
MeOH	4.2 vol	2.0−6.0 vol	4.0−8.0 vol
H_2SO_4	0.3 vol	0.1−0.5 vol	0.05−1.0 vol
trimethyl orthoformate (TMOF)	1.6 vol	0.6−2.0 vol	0.4−1.6 vol
temperature	30 ° C	20−40 ° C	0−40 ° C

Undesired pathway. *Mechanism?*









Reference Point

Design of Experiments: A Merck Case Study

Original Route

Kuethe, Tellers, *Org. Proc. Res. Dev.* **2009**, *13*, 471

5 eq. TsOH
25 eq. TFA
MeCN, 70 °C

Product
50-65%

Both acids required for global deprotection

Improved Alternatives

HTE *1.2g of SM, 236 rxns, 3 days total*

7 eq. H_2SO_4
MeCN/H_2O (90:10), 70 °C
75% Yield
100% Conversion

Product-H_2SO_4

2.2 eq. H_2SO_4
2.2 eq. Octanethiol
MeCN, 20 °C

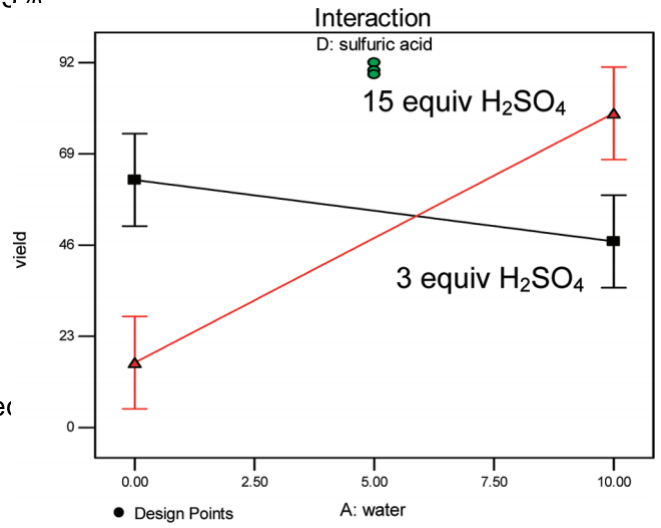
95% isolated yield

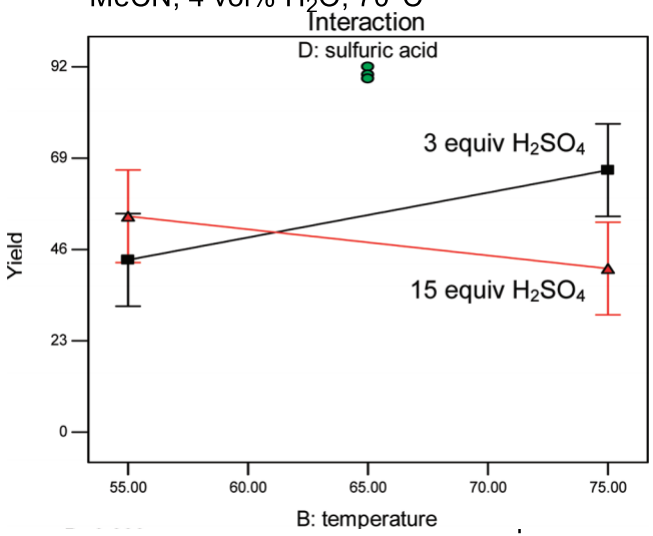
DOE Optimization
19 reactions, central composite

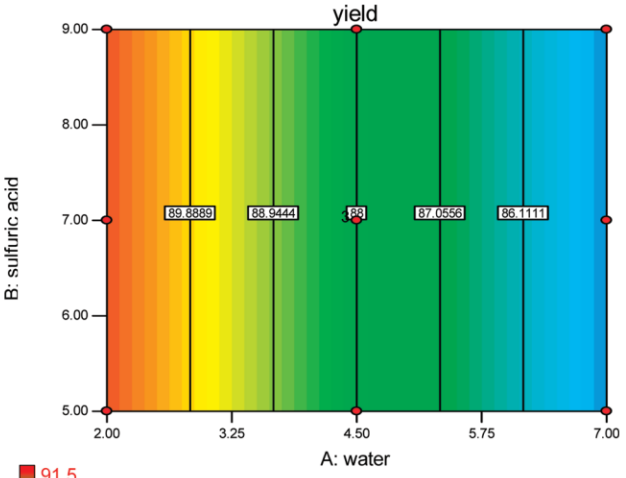
7 eq. H_2SO_4
MeCN, 4 vol% H_2O, 70°C

90% isolated yield

2 H_2SO_4







Key lessons from this work

1) HTS and DoE work together

2) HTS is good for *discrete* variables

3) DoE is great for *continuous* variables

4) DoE, once again, provides valuable insight into variable interactions that guide process decisions

Statistical Modeling of Catalysts by Parameterization

What is the difference between theoretical and empierical models?

Theoretical Models

1) Derived from first principles and do not rely on experimental data
2) All constants have scientific meaning

Emperical Models

1) Require a set of data to fit the model to
2) Mathematical constants do not need to have any meaning

What are some examples of emperical models?

Response surface models (RSM, previously discussed)
Linear Free Energy Relationships (LFERs)

Example of a LFER: The Hammett Equation

$-RTLn(K/K_o) = \Delta G = A/d^2(B_1/D + B_2) \longrightarrow \log(K/K_o) = \sigma\rho = \Delta G$

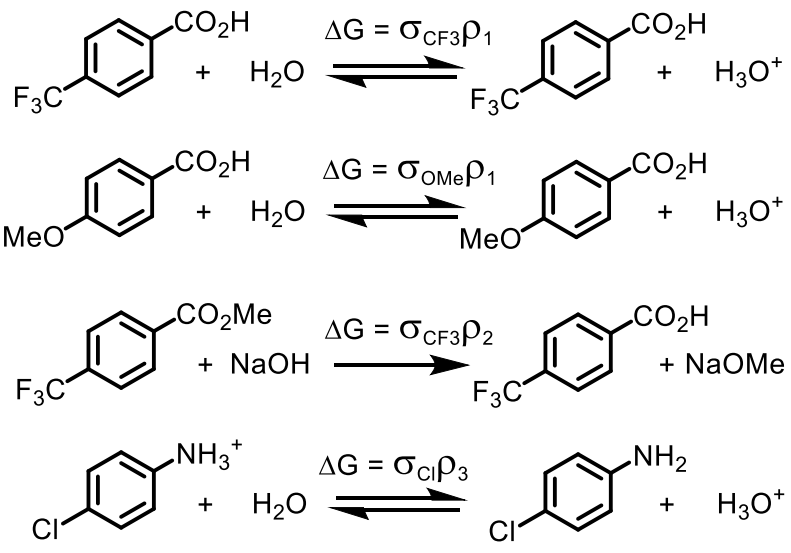
$\sigma = -A/(2.303R)$

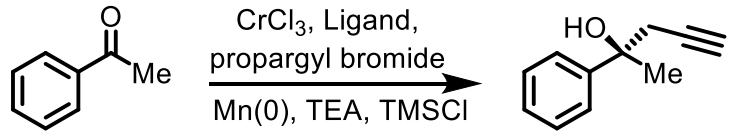
$\rho = (1/Td^2)(B_1/D + B_2)$

A = Constant (changes with subsituent)
B_1 = Constant (changes with reaction)
B_2 = Constant (changes with reaction)
K = rate or equillibrium constant
K_o = reference rate or equillibrium constant
T = Temperature
R = gas constant
D = dielectric constant for solvent
d = distance between the substituent and reaction site

Key Idea

There are parameters which are unique to the substituent and reaction respectively.



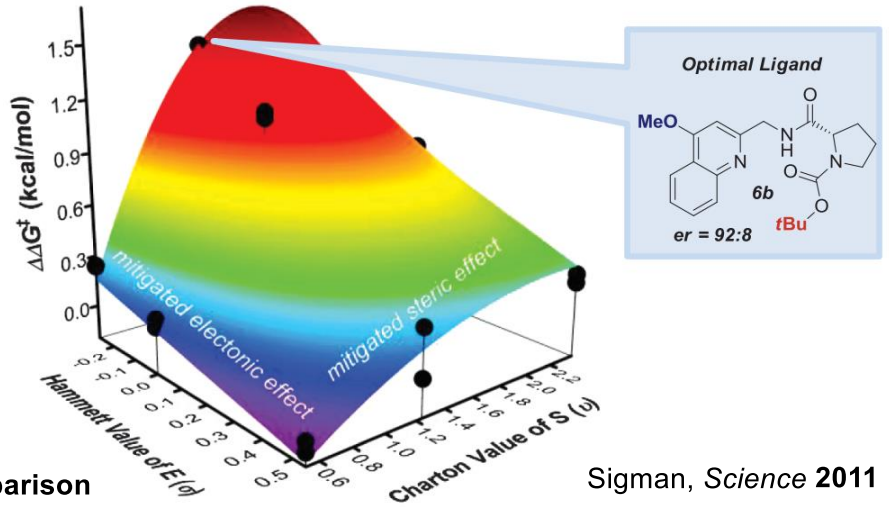


E = Hammett σ Value, S = Charton ν Value

Eq.1) $\Delta\Delta G = -1.20 + 1.22E + 2.84S - 0.85S^2 - 3.79ES + 1.25\,ES^2$

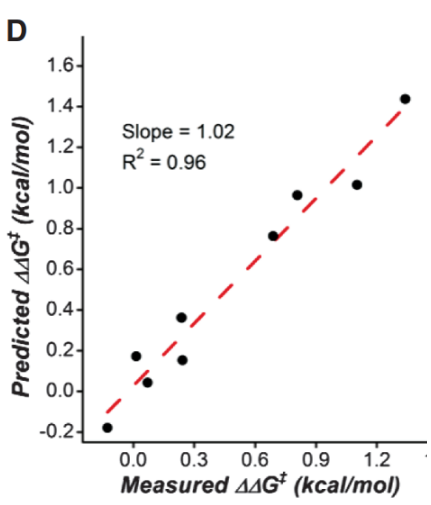
E = Hammett σ Value, B_1 = Minimum Width, B_5 = Maximum Width

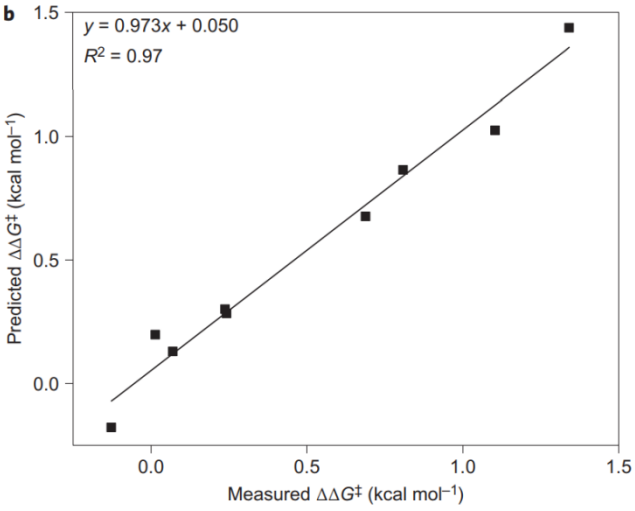
Eq.2) $\Delta\Delta G = -0.696 + 1.380B_1 - 0.962B_5 - 2.705EB_1 + 1.736EB_5$



Sigman, *Science* **2011**

Model Comparison



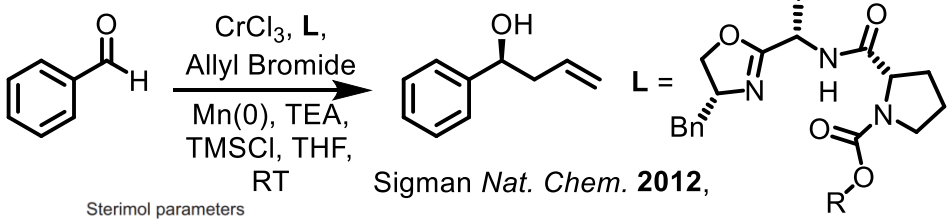


Multivariate LFERs: Parameters

Why is parameterization of organic molecules important?
By nature, molecules and their substituents are discrete variables.

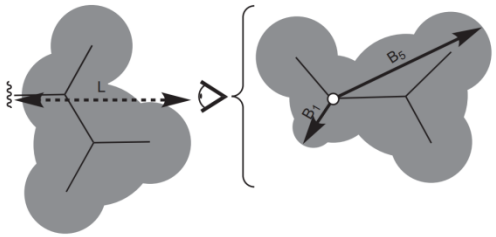
Example: H, Me, *i*-Pr, *t*-Bu, etc.

Discrete variables are useful for *classification*, but not for *regression.* Parameteriation allows us to transform these discrete variables into continuous variables that are valuable for regression. However, this is achieved at the cost of some information and relies on the accuracy with which we can estimate or measure the parameter.

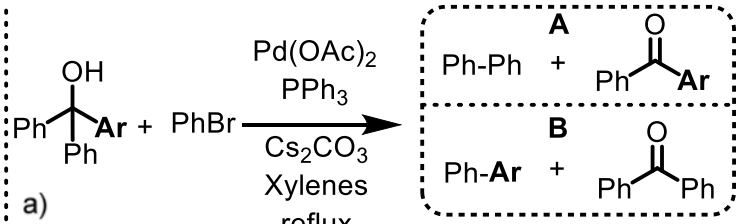


Sigman *Nat. Chem.* **2012**,

Sterimol parameters



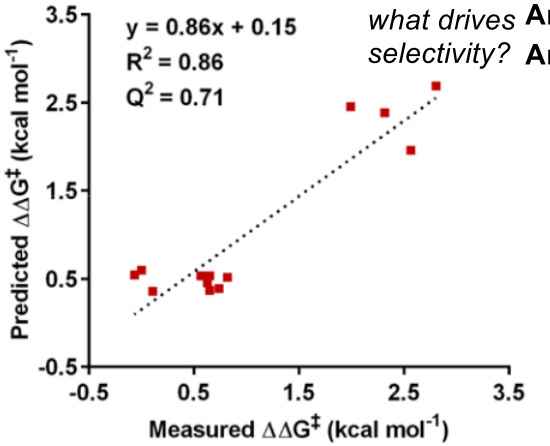
Sterimol paramaters are more information rich than Charton parameters



Hammet correlations were very poor

a)

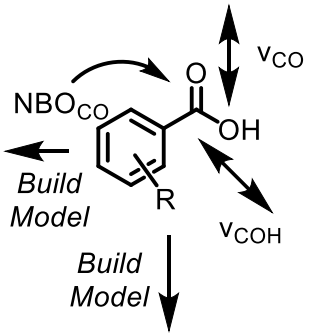
$\Delta\Delta G^{\ddagger} = 1.06 + 0.66NBO_{=O} + 0.58L_o - 0.75\nu_{C=O}$



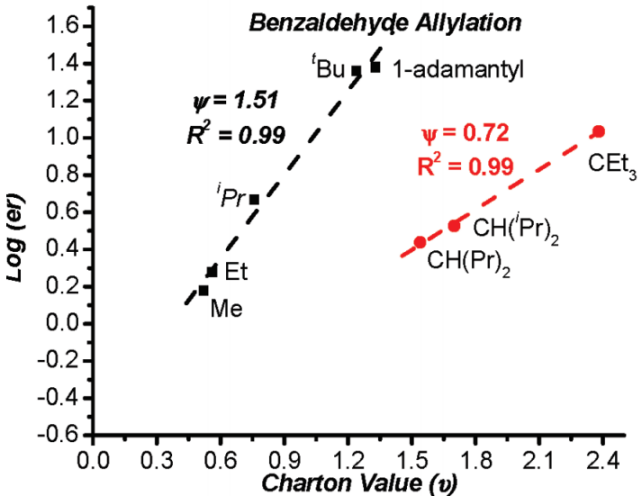
what drives selectivity?

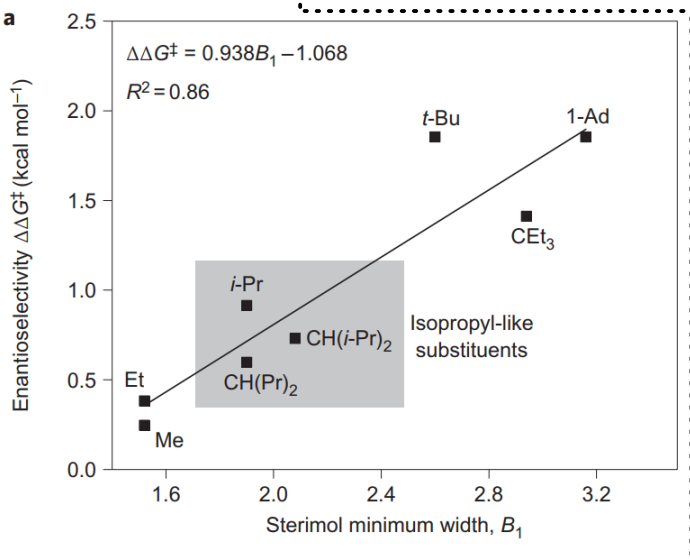
Ar = 2-MeC_6H_4 A:B = 1:26
Ar = 4-MeC_6H_4 A:B = 1:1.2
Ar = 4-$CF_3C_6H_4$ A:B = 1:3

Hammet alternatives?



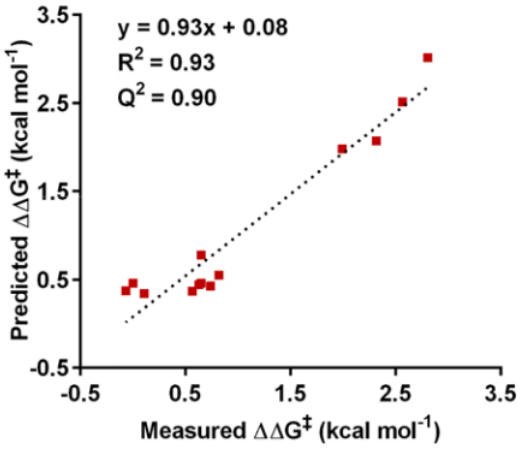
Sigman *J. Am. Chem. Soc.* **2016**, *138*, 13424





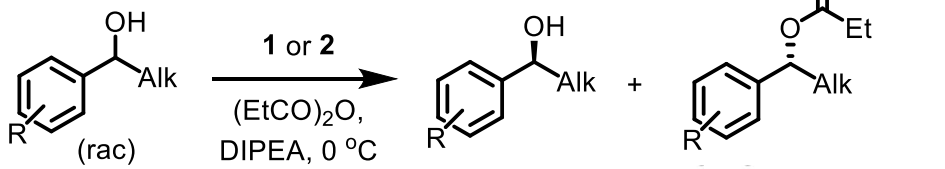
b)

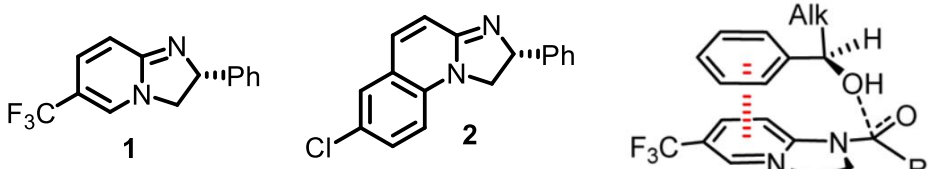
$\Delta\Delta G^{\ddagger} = 1.06 + 0.82B_{1,o} + 0.30\nu_{COH}$



Multivariate LFERs: Mechanism

Whenever an enantioselective reaction requires an aryl group to provide high ee%, people always chalk it up to "pi-pi stacking." How often do you see experimental evidence suggesting this is true?

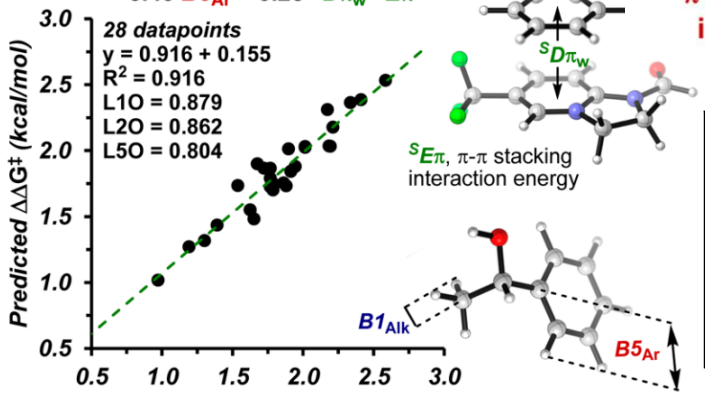




π-π / cation-π interaction

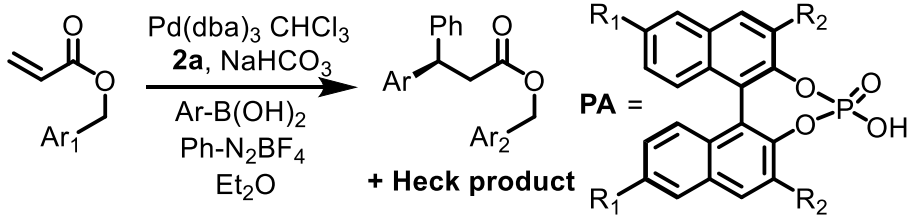
D

$\Delta\Delta G^{\ddagger}$ = - 0.11 - 0.74 $^{S}D\pi_{w}$ + 0.39 $B1_{Alk}$ - 0.40 $B5_{Ar}$ + 0.26 $^{S}D\pi_{w}$ $^{S}E\pi$



The requirement for both the Dπ & Eπ terms indicate that stacking is likely involved in deterimining the ee%

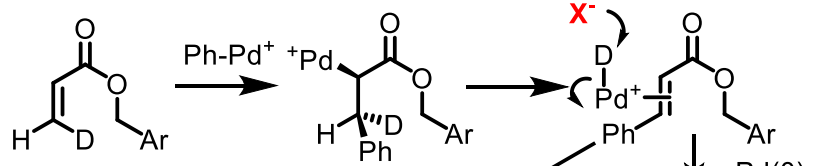
Sigman,Toste, *J. Am. Chem. Soc.* **2016**

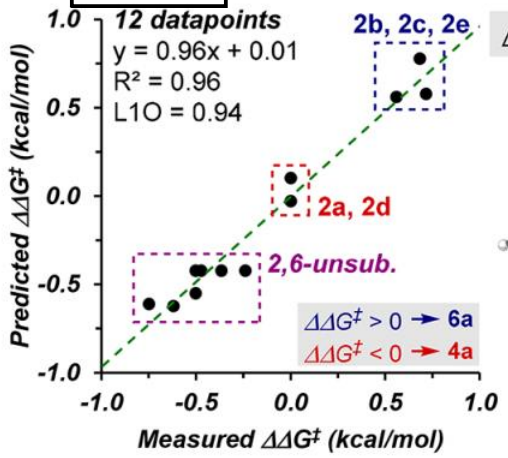


Can multivariate correlations be used to study this reaction?

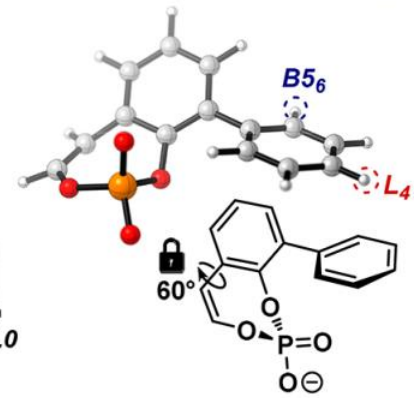
2a: R_1 = C_8H_{17}, R_2 = 9-anthryl

2b: R_1 = H, R_2 = 2,4,6-Cy_3-C_6H_2



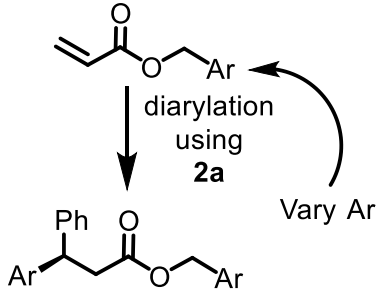


$\Delta\Delta G^{\ddagger}$ = 0.0 - 0.23 L_4 + 1.07 $B5_6$

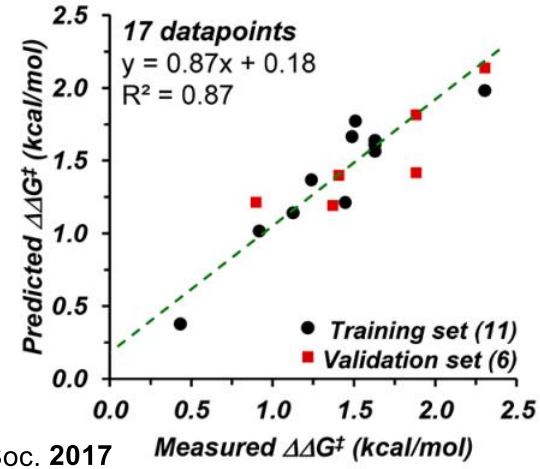


Catalyst sterics can control product ratio suggesting that **2** is involved in the deprotonation event.

What interactions dictate the enantioselectivity?

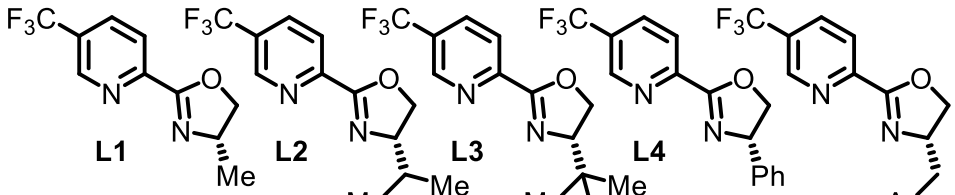


$\Delta\Delta G^{\ddagger}$ = 0.05 + 0.59 NBO_{C1} - 0.42 $^{S}E\pi$ - 0.58 NBO_{C1} $^{S}E\pi$



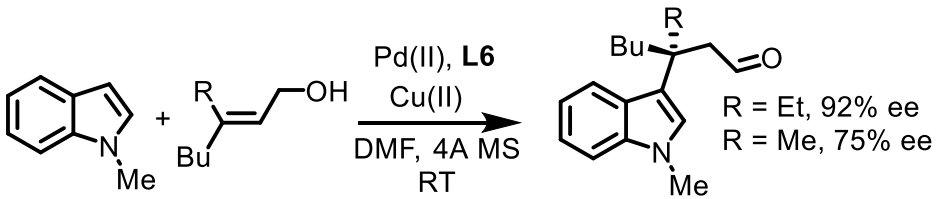
Sigman, Toste, J. Am. Chem. Soc. **2017**

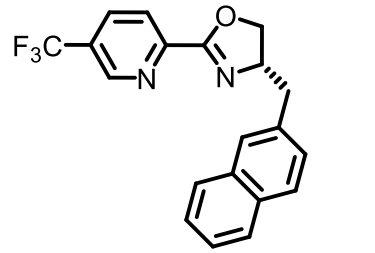
Multivariate LFERs: Prediction



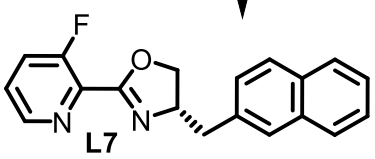
Experimentally evaluated

Sigman, J. Am. Chem. Soc. **2015**, *137*,15668

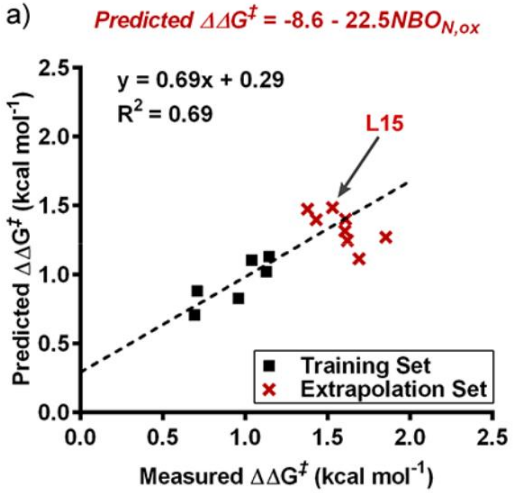




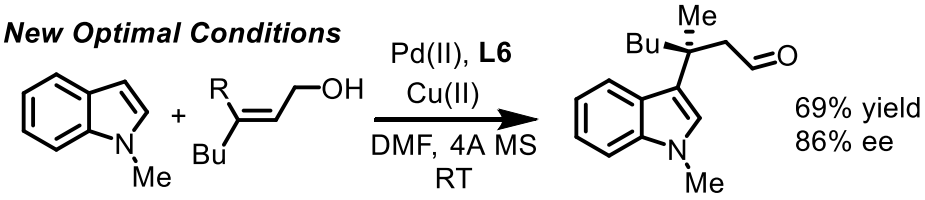
Computationally Screen Ligands



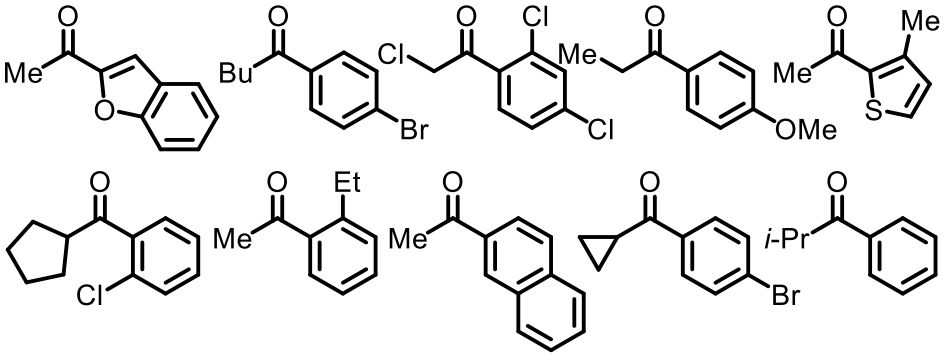
a)



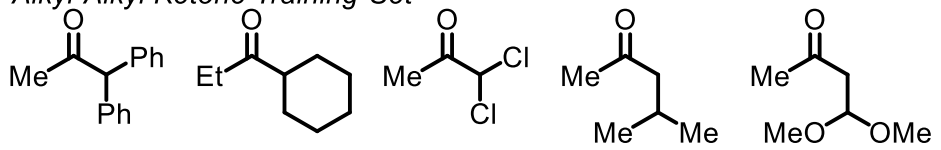
New Optimal Conditions



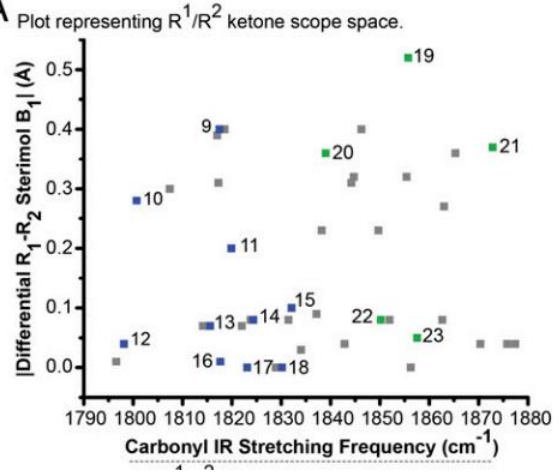
Aryl-Alkyl Ketone Training Set



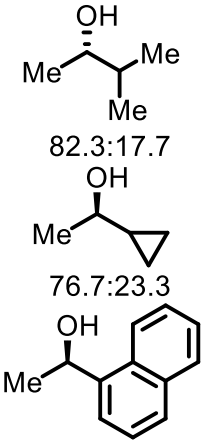
Alkyl-Alkyl Ketone Training Set

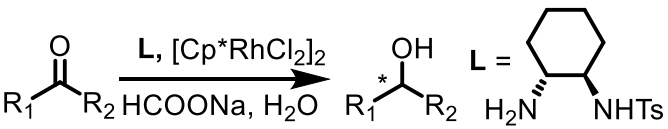


A



Validation Products





< 10% error for these validations

Eq1. (alkyl-alkyl): $\Delta\Delta G = -0.52 - 0.33v_{C=O} - 0.73B_{1s}v_{sci} - 1.75v_{sci}I_{sci}$

Eq2.: $\Delta\Delta G = -0.10 - 0.38v_{C=O} - 0.8B_1 + 0.29B_{1Ar} - 0.52Tor - 1.46v_{C=O}B_1$

Brief Timeline of Computer Aided Synthesis Efforts

For a more thorough review, see Maimone GM "Computer-Assisted Organic Synthesis"

1969
E. J. Corey Publishes first paper on CAR

1977
ACS published a symposium series titled "Computer-Assisted Organic Synthesis"

E.J. Corey's contribution to this symposium was to detail "Logic and Heuristics Applied to Synthetic Analysis" (LHASA)

1990
William Jorgensen, develops a "Computer Assisted Mechanistic Evaluation of Organic Reactions" (CAMEO)

1990
Herbery Gelertner introduces machine learning into SYNCHEM.

This marks the first use of machine learning in organic chemistry.

Current State of the Art
(next few slides)

Future Prospects
A fully autonomous CAR?

Why am I lumping machine learning into a statistics group meeting?

Because the topics are related and in terms of their application to organic synthesis, they are very similar.

What are the primary problems machine learning tries to solve?

Classification: automatedly learning what something is and what category it belongs to.

Regression: automatedly building mathematical models that describe the relationship of variables to one another.

What are the primary kinds of machine learning?

Supervised: learning from a dataset that has answers.

Unsupervised: learning patterns in a dataset without being explicitly told what to look for.

What are some popular supervised machine learning algorithms?

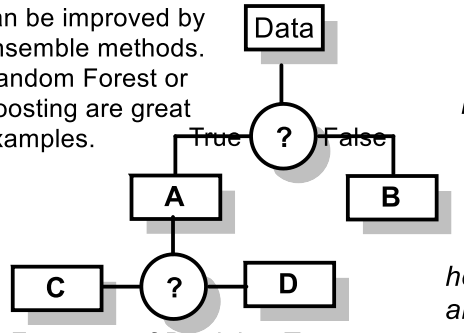
Note: There are many more, these are just popular and relevant to today's topic.

Generalized Linear Models
- Multiple Linear Regression (previously discussed)
- Logistic Regression (used for obtaining odds ratios)

Features of GLMs: Quick and easy to train. Occam's razor of sorts. The ability to interpret their decisions is inverseley proportional to the number of variables used.

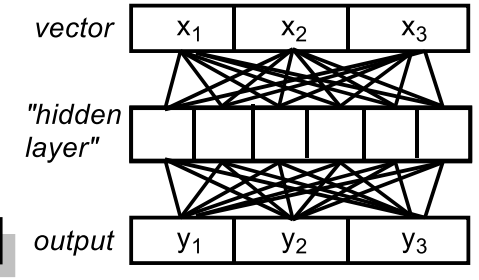
Note: decision trees can be improved by ensemble methods. Random Forest or Boosting are great examples.

Decision Tree



Features of Decision Trees:
Extremely easy to interpret.
Exremely fast to train.

Artificial Neural Network



how to think about NNs **Black Box**

NN features: Powerful, yet difficult to interpret decision making.

Machine Learning in Synthetic Chemistry

An interesting idea

Can a computer learn frontier molecular orbital theory?

Reaction Explorer
Hand-coded rule based system
1,500 rules

Baldi, *J. Chem. Inf. Model.* **2009**, *49*, 2034

Train a NN to find patterns in "sinks" and "sources" given frontier MOs of all reactants and example reactions.

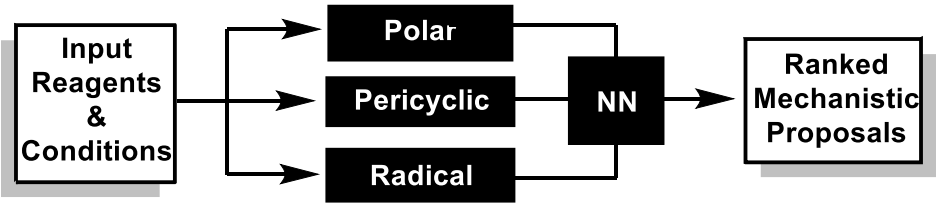
Prototype ReactionPredictor
Given "filled" and "unfilled" MOs, can predict polar mechanisms

Same idea, but include radicals and pericyclics

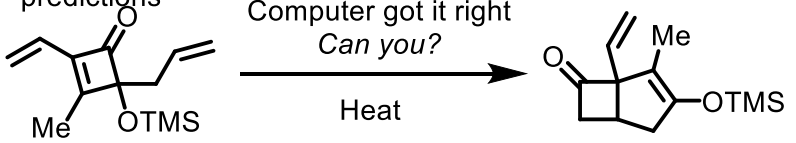
ReactionPredictor

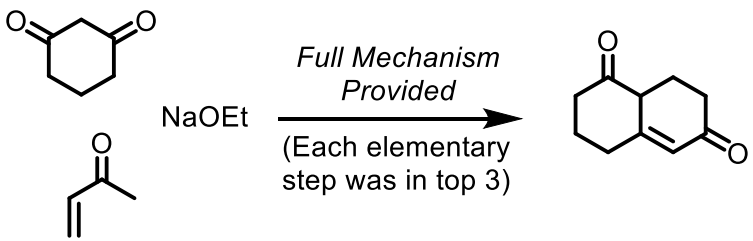
Baldi, *J. Chem. Inf. Model.* **2011**, *51*, 2209

How does reaction predictor work?



> 77% accuracy >93% correct mech. in the top 4 predictions





Baldi, *J. Chem. Inf. Model.* **2012**, *52*, 2526

Reaction Classifier

Landrum, *J. Chem. Inf. Model.* **2015**, *55*, 39

Patent Literature reactions
1,109,897

NameRxn
NextMove Software
(hand coded system)

Classified Reactions (54% of initial set)
599,344

Filter by top 50 most frequent reactions

Note: 200 reactions were randomly selected from each category for training and 800 reactions were slected randomly from each category for testing. Various fingerprinting techniques were also evaluated at this stage.

Highest Class
Carboxylic Acid+Amine
44,675
48 other reactions
Nitrile Reduction
2,662
50th class

Split & Train
(5 different ML models evaluated)

Optimal Machine Learning Models
Logistic Regression & Random Forest

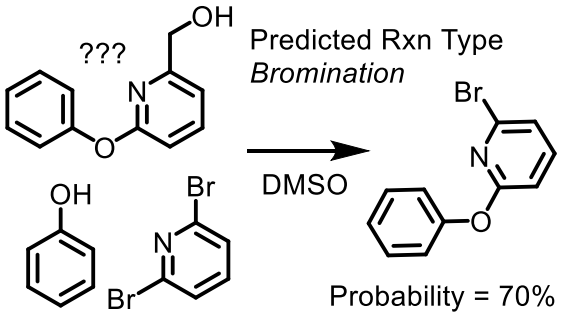
(Logistic regression model)

Apply predictions to electronic notebook data (38,326 entries)

Precision = 86%
Recall = 93%

Note: this program can classify reactions that NameRxn cannot

Problem with patent data



Notable Reaction Classifications

Stille
2,796

Sonogashira
7,071

Bromo Suzuki
Bromo Suzuki-Type
Chloro Suzuki
17,759

Fluoro N-Arylation
Chloro N-Arylation
Fluoro N-Arylation
35,987

Machine Learning in Synthetic Chemistry

Ranking predictor workflow Jensen, *ACS Cent. Sci.* **2017**, *3*, 434

Reactions from patents between 1976 & 2013

1,122,662

Heuristic-driven template extraction

Extracted Templates

140,284

reduce templates to those with >50 examples

Filtered Templates

1,689

Apply templates on 15,000 examples

"Possible Reactions"

5,335,669

Split Data

Data Splits

70% Training
10% Validation
20% Testing

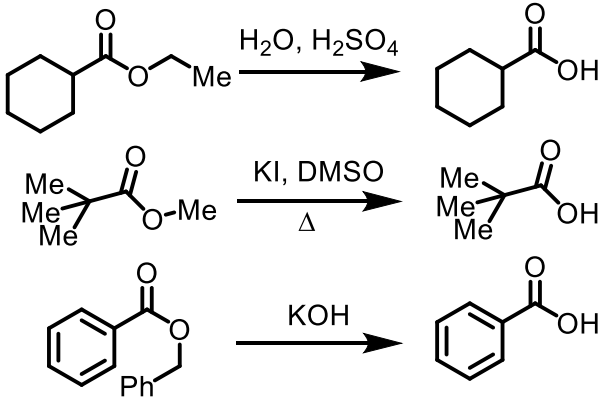
Train Neural Network

5-fold cross validation
leave out "Validation"

Neural Network Model

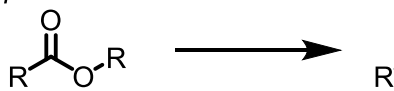
What is a template?

Observations



Heuristic-driven template extraction (Algorithm)

Template



Validation Data → **Neural Network Model**

Accurately Predicted Major Product
68.5%
(average undergrad?)

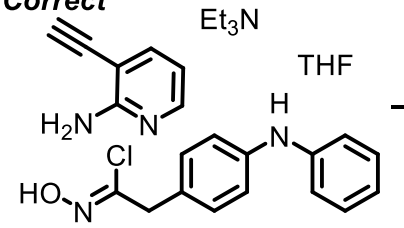
Major product identified in top **X** predictions
X = 3, 84.8%
X = 5, 89.4%

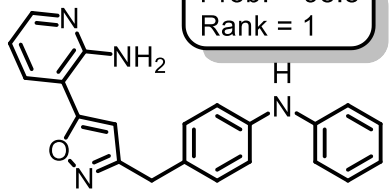
Note 1: while the validation data was not used during the model training, the data used for this protocol was limited to reactions where at least 1 of the templates could provide the desired product.

Note 2: When the 1,689 templates used in this study were applied to a random set of 15,000 reactions, the actual product was found in 76% of examples. This arises from the filtering of templates with less than 50 reported examples.

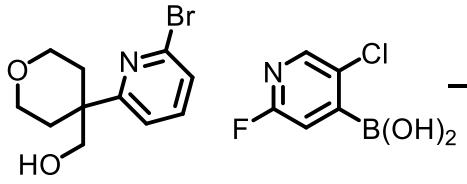
Examples of Predicted Reactions
(no reagents are missing, patent data isn't clean)

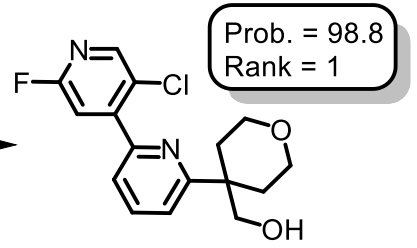
Correct



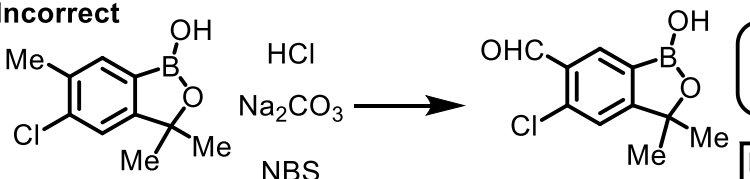


Correct





Incorrect



Note: Rank 1 probability was only 26.9%

Machine Learning in Synthetic Chemistry

Filter reactions by:
A + B + C → D

Reaxys → (undisclosed data gathering algorithm) → **Undisclosed Sample Size** (millions)

Learn rules based on bond changes and neighboring atoms

A) 17,370 rules
B) 8,720 rules
C) 137 rules

save rules if they occur
A) >50 times
B) >100 times
C> >5,000 times

Filter by rules

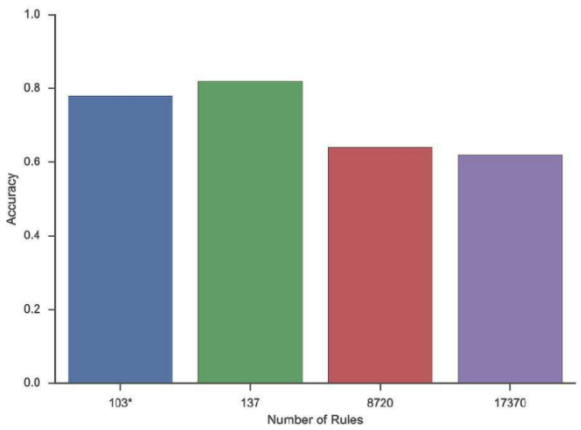
Filter by 103 hand written rules by an organic chemist

Split and train datasets using 5-fold cross-validation 70:20:10 Train:Test:Validate

Reactions **4,900,000** → **Artificial Neural Network**; **Logistic Regression Model**

Reactions **3,000,000** → **Logistic Regression Model**; **Artificial Neural Network**

a) Influence of the rule set size

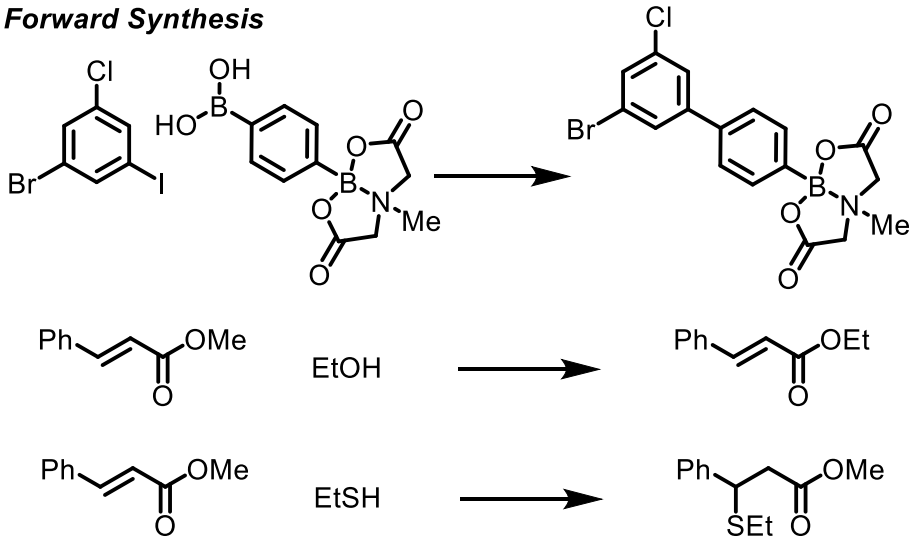


Accuracy Metrics

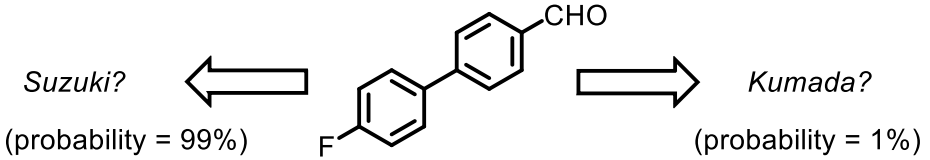
Retrosynthesis
Accuracy LR: 31%
Accuracy NN: 62%
MRR LR: 0.41
MRR NN: 0.75

Forward
Accuracy LR: 41%
Accuracy NN: 77%
MRR LR: 0.49
MRR NN: 0.85

Forward Synthesis

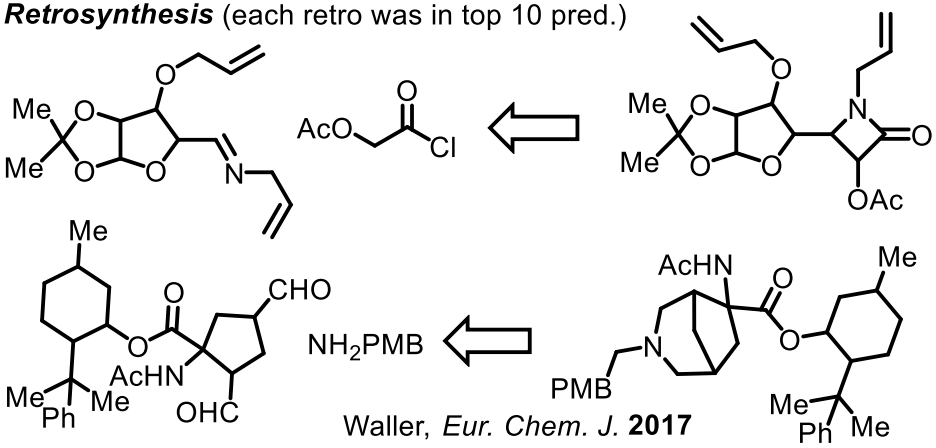


Interesting Observation



This selectivity was learned, not programmed

Retrosynthesis (each retro was in top 10 pred.)



Waller, *Eur. Chem. J.* **2017**