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 determine 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?

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 occurrences, 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 relationship 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/](https://onlinecourses.science.psu.edu/statprogram/)
Types of Experiments

One Variable at a Time (OVAT)

Hold X & Y Constant

Vary X

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.

Vary Y

Pick optimal X

Z Constant

Pick optimal Y

Vary Z

Factorial Design (DoE)

Vary X, Y Z simultaneously

use mapping to predict optimal conditions

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 its curvature.

Central Composite Design

most popular response surface methodology

Apply 2$^F$ "star points"

The center points are commonly replicated N times so that you can estimate 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

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

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

---

**Table**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Initial set-points (half-factorial study)</th>
<th>Ranges (RSM study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeOH</td>
<td>4.2 vol</td>
<td>2.0–6.0 vol</td>
</tr>
<tr>
<td>H₂SO₄</td>
<td>0.3 vol</td>
<td>0.1–0.5 vol</td>
</tr>
<tr>
<td>Trimethyl orthoformate (TMOF)</td>
<td>1.6 vol</td>
<td>0.6–2.0 vol</td>
</tr>
<tr>
<td>Temperature</td>
<td>30 °C</td>
<td>20–40 °C</td>
</tr>
</tbody>
</table>

---

**Undesired pathway. Mechanism?**

---

**Graphs**

- **Product (3)**: 77.75
- **Pleuromutilin (2)**: 5.45
- **Alkene (14)**: 4.50739

---

**Equations**

- **Equation 1**: \( \text{Reaction} \rightarrow \text{Product} \)
- **Equation 2**: \( \text{Mechanism} \)

---

**Graphs**

- **Product (3)**: 77.75
- **Pleuromutilin (2)**: 5.45
- **Alkene (14)**: 4.50739
Design of Experiments: A Merck Case Study

Original Route

Me Me N H
O O CN
Me Me N H

Improved Alternatives

Me Me N H
O O CN
Me Me N H


Product

50-65%

Both acids required for global deprotection

HTE

1.2g of SM
236 rxns
3 days total

7 eq. H₂SO₄

MeCN/H₂O (90:10), 70 °C
75% Yield
100% Conversion

Product-H₂SO₄

DOE Optimization

19 reactions, central composite

90% isolated yield

MeCN, 4 vol% H₂O, 70°C

Interaction

D: sulfuric acid

15 equiv H₂SO₄

3 equiv H₂SO₄

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 empirical models?

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

**Empirical 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 empirical models?

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

**Example of a LFER: The Hammett Equation**

\[-RT\ln(K/K_0) = \Delta G = A/d^2(B_1/D + B_2)\]

- \(A\) = Constant (changes with substituent)
- \(B_1\) = Constant (changes with reaction)
- \(B_2\) = Constant (changes with reaction)
- \(K\) = rate or equilibrium constant
- \(K_0\) = reference rate rate or equilibrium constant
- \(T\) = Temperature
- \(R\) = gas constant
- \(D\) = dielectric constant for solvent
- \(d\) = distance between the substituent and reaction site

\[\log(K/K_0) = \sigma \rho = \Delta G\]

\[\sigma = -A(2.303R)\]

\[\rho = \frac{1}{2}(B_1/D + B_2)\]

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

**Model Comparison**

**Sigman, Science 2011**

\[\text{Slope} = 1.02\quad R^2 = 0.96\]

\[\text{Slope} = 0.97\quad R^2 = 0.97\]
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. Parameterization 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.

Sterimol parameters are more information rich than Charton parameters.

Sigman *Nat. Chem.* 2012,

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?

\[
\begin{align*}
\text{(rac)} & \quad \text{1 or 2} \quad \text{(EICO)}_2\text{O}, \ DIPEA, \ 0^\circ \text{C} \quad \text{Ph-Pd}^{+} \quad \text{Pd}^{+} \\
\text{Alk} & \quad \text{OH} \quad \text{Alk} \quad \text{EtO} \\
\text{R} & \quad \text{F}_{3}\text{C} \quad \text{Ph} \quad \text{Cl} \\
\text{1} & \quad \text{2} \\
\end{align*}
\]

\[\Delta \Delta G^\ddagger = -0.11 - 0.74 \, S_{D\pi w} + 0.39 \, B_{1Alk} - 0.40 \, B_{5Ar} + 0.26 \, S_{D\pi w} \, E_{\pi} \]

\[12 \text{ datapoints} \quad y = 0.96x + 0.01 \quad R^2 = 0.96 \quad L10 = 0.94 \]

\[2b, 2c, 2e \quad \text{insertion} \quad \text{Heck Product} \]

\[\Delta \Delta G^\ddagger = 0.0 - 0.23 \, L_4 + 1.07 \, B_{56} \]

\[28 \text{ datapoints} \quad y = 0.916 + 0.155 \quad R^2 = 0.879 \quad L10 = 0.862 \quad L50 = 0.804 \]

\[\text{Predicted } \Delta \Delta G^\ddagger (\text{kcal/mol}) \quad \text{Measured } \Delta \Delta G^\ddagger (\text{kcal/mol}) \quad \text{Sigman, Toste, J. Am. Chem. Soc. 2016} \]

\[\text{2a: } R_1 = C_8H_{17}, \ R_2 = 9\text{-anthryl} \quad 2b: \ R_1 = H, \ R_2 = 2,4,6-Cy_3-C_6H_2 \]

\[\text{Catalyst sterics can control product ratio suggesting that 2 is involved in the deprotonation event.} \]

\[\text{What interactions dictate the enantioselectivity?} \]

\[\Delta \Delta G^\ddagger = 0.05 + 0.59 \, NBO_{C1} - 0.42 \quad S_{E\pi} = 0.58 \, NBO_{C1} \quad S_{E\pi} \]

\[17 \text{ datapoints} \quad y = 0.87x + 0.18 \quad R^2 = 0.87 \]

\[\text{Can multivariate correlations be used to study this reaction?} \]

\[\text{Sigman, Toste, J. Am. Chem. Soc. 2017} \]

\[\text{PA} = \begin{align*}
\text{Ar}_1 & \quad \text{Ph} \\
\text{R}_1 & \quad \text{R}_2 \\
\text{Ar}_2 & \quad \text{PA} \\
\text{Ar} & \quad \text{Ar} \\
\end{align*} \]

\[\text{diarylation using 2a} \quad \text{Vary Ar} \]
Multivariate LFERs: Prediction

F₃C- \( \text{N} \) \( \text{O} \) \( \text{N} \) \( \text{Me} \) \( \text{L1} \) 
F₃C- \( \text{N} \) \( \text{O} \) \( \text{N} \) \( \text{Me} \) \( \text{L2} \) 
F₃C- \( \text{N} \) \( \text{O} \) \( \text{N} \) \( \text{Me} \) \( \text{L3} \) 
F₃C- \( \text{N} \) \( \text{O} \) \( \text{N} \) \( \text{Me} \) \( \text{L4} \) 

Experimentally evaluated
Sigman, J. Am. Chem. Soc. 2015, 137, 15668

\[ \text{L5 \( \text{Ar} = \text{Ph} \) } \]
\[ \text{L6 \( \text{Ar} = 2-\text{Np} \) } \]

\( \text{N} \) \( \text{Me} \) \( \text{Bu} \) \( \text{R} \) \( \text{OH} \) \( \text{Pd(II), L6} \) \( \text{Cu(II)} \) \( \text{DMF, 4A MS} \) \( \text{RT} \)

\[ \text{C} \] \( \text{R} \) \( \text{N} \) \( \text{Me} \) \( \text{Bu} \) \( \text{R} \) \( \text{OH} \) \( \text{Pd(II), L6} \) \( \text{Cu(II)} \) \( \text{DMF, 4A MS} \) \( \text{RT} \)

\[ \text{Predicted } \Delta \Delta G^\ddagger = -8.6 - 22.5 NBO \_ {\text{N,ox}} \]

\[ y = 0.69x + 0.29 \]
\[ R^2 = 0.69 \]

\[ \text{Plot representing } R_1^2/R_2^2 \text{ ketone scope space.} \]

\[ \text{Differential } R^1 \_ {\text{R}2} \text{ Sterol } B_1(\text{A}) \]

\[ \text{Eq1. (alkyl-alkyl): } \Delta \Delta G = -0.52 - 0.33v_{\text{C=O}} - 0.73B_{\text{1s}}v_{\text{sci}} - 1.75v_{\text{sci}}^2 \]
\[ \text{Eq2.: } \Delta \Delta G = -0.10 - 0.38v_{\text{C=O}} - 0.8B_{\text{1s}} + 0.29B_{\text{1Ar}} - 0.52\text{Tor} - 1.46v_{\text{C=O}}B_{\text{1s}} \]

\[ \text{Validation Products} \]

\[ \text{OH} \]
\[ \text{Me} \_ {\text{Me}} \]
\[ \text{Me} \_ {\text{Me}} \]
\[ \text{Me} \_ {\text{Me}} \]
\[ \text{OH} \]
\[ \text{Me} \]
\[ \text{Me} \]
\[ \text{OH} \]
\[ \text{Me} \]
\[ \text{88.4:11.6} \]

\[ \text{OH} \]
\[ \text{Me} \]
\[ \text{Me} \]
\[ \text{OH} \]
\[ \text{Me} \]
\[ \text{82.3:17.7} \]

\[ \text{OH} \]
\[ \text{Me} \]
\[ \text{Me} \]
\[ \text{OH} \]
\[ \text{Me} \]
\[ \text{76.7:23.3} \]
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.

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

Artificial Neural Network

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

Black Box

NN features: Powerful, yet difficult to interpret decision making.
An interesting idea
Can a computer learn frontier molecular orbital theory?

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


Same idea, but include radicals and pericyclics

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


How does reaction predictor work?

**Input Reagents & Conditions**

Polar
Pericyclic
Radical

**NN**

Ranked Mechanistic Proposals

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

Computer got it right Can you?

Heat

**Full Mechanism Provided**
(Each elementary step was in top 3)

Oxetane

NaOEt

OAc

Me

Heat

**Problem with patent data**

Predicted Rxn Type Bromination

Bromo Suzuki
Chloro Suzuki
Fluoro N-Arylation
Fluoro N-Arylation

Stille
Sonogashira
Bromo Suzuki-Type Chloro Suzuki
Fluoro N-Arylation
Chloro N-Arylation
Fluoro N-Arylation


**Reaction Classifier**

Patent Literature reactions

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

**NameRxn**

**NextMove Software**

(hand coded system)

Classified Reactions
(54% of initial set)

Note: 200 reactions were randomly selected from each category for training and 800 reactions were selected 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

**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

**Split & Train**

(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

**Filter by top 50 most frequent reactions**
Machine Learning in Synthetic Chemistry

Ranking predictor workflow

Jensen, ACS Cent. Sci. 2017, 3, 434
Reactions from patents between 1976 & 2013

Heuristic-driven template extraction

1,122,662

Extracted Templates

140,284

What is a template?

Observations

\[
\begin{align*}
\text{O} & \quad \text{H}_2\text{O}, \text{H}_2\text{SO}_4 \\
\text{O} & \quad \text{KI}, \text{DMSO} \\
\text{Ph} & \quad \text{KOH}
\end{align*}
\]

Filtered Templates

1,689

Heuristic-driven template extraction (Algorithm)

"Possible Reactions"

5,335,669

Neural Network Model

Train Neural Network

5-fold cross validation leave out “Validation”

70% Training
10% Validation
20% Testing

Neural Network Model

Validation Data

Accurately Predicted
Major Product

68.5%
(average undergrad?)

Major product identified in top X predictions

\[ X = 3, \ 84.8\% \]

\[ X = 5, \ 89.4\% \]

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

Correct

Prob. = 98.8
Rank = 1

Correct

Prob. = 98.8
Rank = 1

Incorrect

Prob. = 10^{-5}
Rank = 190

Note: Rank 1 probability was only 26.9%
Machine Learning in Synthetic Chemistry

**Reaxys**
Filter reactions by:
A + B + C → D
(undisclosed data gathering algorithm)

**Undisclosed Sample Size**
(millions)

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

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

Learn rules based on bond changes and neighboring atoms

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**

**Logistic Regression Model**

**Artificial Neural Network**

Reactions
3,000,000

**Forward Synthesis**

**Interesting Observation**
Suzuki? (probability = 99%)
Kumada? (probability = 1%)
*This selectivity was learned, not programmed*

**Retrosynthesis** (each retro was in top 10 pred.)

**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

Waller, Eur. Chem. J. 2017