DISCLAIMER - Dilletennte - a person who cultivates an area of interest, such as the arts, without real commitment or knowledge

High Throuput Experimentation (HTE) is transforming chemistry, drug discovery, and biology

Machine Learning has displayed equally disruptive effects in biology and medicine but progress in chemistry has not been commensurate with these respective advances

"Field of study that gives computers the ability to learn without being explicitly programmed" - Arthur Samuel 1959

Data is critically needed to enable machine learning, thus empowered by HTE

Included Topics
- History of development
- Common types of learning methods
- Organic methodology
- Chemical synthesis
- Drug design
- Future outlooks from the literature

Topics Not Included
- HTE besides selected miniaturization
- Heterogenous catalysis
- Materials or property prediction
- Binary Classification ML

For popular science introductions see:
- Second Machine Age - Brynjolfsson and McAfee
- Life 3.0 - Max Tegmark
- Superintelligence - Bostrom

Useful Technical References
- Artificial Intelligence - Norwig and Russell
- Elements of Statistical Learning - Friedman, Tibshirani, Hastie

In my view, those efforts are certainly worthwhile as long as they don’t come at the expense of the very field they wish to simplify. In other words, such efforts do not intimidate, threaten, or provoke fear in the hearts of any practitioner of synthesis. Promises of computational chemistry and combinatorial chemistry displacing the field were made over the years, yet we are still here. - Baran 2018
1642 - Pascal invents first mechanical calculator

1828 - Wöhler synthesis of urea

1666-1674 - Newton and Leibniz begin laying foundation for modern calculus

1845 - Boole sets out to explore symbolic representations of reasoning, inventing Boolean Algebra

1913 - Principia Mathematica by Russell, is published, exploring Formal Logic

1931 - Gödel’s incompleteness theorems

1932 - Sherrington shares Nobel Prize for contributions in neuroscience - discovers role of ‘synapse’

1944 - McCulloch and Pitts develop algorithm for Artificial Neural Networks (ANN)

1944 - Woodward and Doering complete the total synthesis of quinine - representing first ‘modern’ total synthesis

1986 - Hinton reports backpropagation method enabling practical ANN

1997 - Google Deepblue beats world chess master Kasparov 3.5-2.5

2003 - Human genome project is declared complete, largest collaborative biology project in world history

2010 - Nobel Prize in chemistry for palladium-catalyzed cross couplings

2011 - IBM Watson beats Jeopardy champions

2015 - AlphaGo becomes first computer to beat a human player without handicaps in Go
<table>
<thead>
<tr>
<th><strong>Algorithm</strong></th>
<th>Process or set of rules to be followed in calculations or other problem solving operations, especially by a computer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Artificial Intelligence</strong></td>
<td>Intelligence demonstrated by machines</td>
</tr>
<tr>
<td><strong>Artificial Neural Networks</strong></td>
<td>Inspired by biological neural networks - large number of inputs into a targeted output</td>
</tr>
<tr>
<td><strong>Backpropagation</strong></td>
<td>Error attribution of coefficient weights from output layer, back to input layer</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>Qualitative data output - cat vs. dog image recognition</td>
</tr>
<tr>
<td><strong>Deep Learning</strong></td>
<td>Machine learning algorithm based on multilayer neural networks</td>
</tr>
<tr>
<td>Uses a series of non-linear functions combinatorially, creating multiple levels of representations for different levels of abstraction</td>
<td></td>
</tr>
<tr>
<td><strong>Go</strong></td>
<td>One of the world's most complex games, $10^{170}$ possible positions</td>
</tr>
<tr>
<td><strong>Hueritics</strong></td>
<td>An approach to problem solving, learning, or discovery that employs a practical method not guaranteed to be optimal or perfect, but likely sufficient</td>
</tr>
<tr>
<td><strong>High-Throughput-Experimentation</strong></td>
<td>Experimentation that provides large amounts of data that far exceeds traditional lab-scale preparations; examples include DEL (DNA-Encoded-Libraries), maturation, and on-chip</td>
</tr>
<tr>
<td><strong>Least-squares linear regression</strong></td>
<td>Simplest machine learning algorithm (Hammett Plot)</td>
</tr>
<tr>
<td><strong>Machine Learning</strong></td>
<td>Study and construction of computer algorithms that can learn from data</td>
</tr>
<tr>
<td><strong>Molecular descriptors</strong></td>
<td>Describes properties of any compound by utilizing domain expertise</td>
</tr>
<tr>
<td><strong>Monte Carlo Tree Search</strong></td>
<td>Heuristic search algorithm for some kinds of decision processes, most notably those employed in game play</td>
</tr>
<tr>
<td><strong>Neuron</strong></td>
<td>Biological cell-type that transmits information in the nervous system</td>
</tr>
<tr>
<td><strong>Processing Unit</strong></td>
<td>Maps features into an output using a function - simplest unit of deep learning algorithms; artificial neuron</td>
</tr>
<tr>
<td><strong>QSAR</strong></td>
<td>Quantitative Structure Activity Relationship - goal to predict the biological activity of a certain compound</td>
</tr>
<tr>
<td><strong>Random Forest Model</strong></td>
<td>Machine learning algorithm that takes random samples of various decision trees, and combines their inputs to select an output</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>Quantitative data output - reaction yield</td>
</tr>
<tr>
<td><strong>Reinforcement Learning</strong></td>
<td>Machine learning paradigm between supervised and unsupervised; data is not labelled, but correct/incorrect response is made</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>Root Mean Squared Error - Difference between values predicted and values observed; $R^2$ - relative error</td>
</tr>
<tr>
<td><strong>Shallow ML</strong></td>
<td>Input features provided by domain expert; template matching (does not learn representation of problem)</td>
</tr>
<tr>
<td><strong>Support Vector Machine</strong></td>
<td>Machine learning algorithm best for classification processes; utilizes 'kernels' to find boundaries between different classes</td>
</tr>
<tr>
<td><strong>Supervised Learning</strong></td>
<td>Paradigm in which input has a labelled output for training set; reaction parameters as input, reported yields as output</td>
</tr>
<tr>
<td><strong>Test Set</strong></td>
<td>Set of data that is utilized to generate a predictive algorithm, usually exploring many types of methods, i.e. linear regression, ANN, Bayes classifier, or Random Forests; train algorithm on 70% of experiments, test on the other 30%</td>
</tr>
<tr>
<td><strong>Training Set</strong></td>
<td>Set of data that is reserved to test the predictive ability of the algorithm; train algorithm on 70% of experiments, test on the other 30%</td>
</tr>
<tr>
<td><strong>Unsupervised Learning</strong></td>
<td>Paradigm where the inputs are completely unlabelled, algorithm itself comes up with classifications to attach to the data and arrives independently at outputs</td>
</tr>
</tbody>
</table>
Techniques for analyzing data set - choose based on specific problem

Similar to computational chemistry - need to choose the correct level of theory based on the problem at hand

**Decision Trees**

CART - Classification and Regression Trees
- Node - test of an attribute
- Branch - outcome of test
- Leaf - class label
- Low data curation
- Overfitting is primary problem
- Grow tree - Choose features
  - Identify splitting conditions
  - Learn when to stop
- Prune tree
- **Applications** - Titanic survival statistics, male or female based on height and weight, price of home

**Random Forest**

- Grow multiple trees
- More robust and accurate with multiple models
- For a new object, each tree gives a vote
- Output with the most 'votes' gets chosen
- Avoids overfitting problem of single decision tree
- Large data sets with high dimensionality
- Disadvantage - black box
- **Applications** - banking, epidemiology, stock behavior, reaction yield, voice classification

**Artificial Neural Networks (ANN)**

- Biological inspiration
- Artificial neuron - perceptron
  - Multiple inputs to generate one output
  - Each neuron is weighted
  - Distinct activation functions to decide output
  - Unit step, sign, linear, logistic, etc...
  - Multiple types of outputs
- Train either forwards or backwards
- **Applications** - self-driving cars, image recognition

**Deep Learning**

- ANN with multiple hidden layers
- Computationally intractable until recently
- Made feasible with rectified linear unit activation - transformative impact
- Gradient descent for error correction
- **Applications** - AlphaGo, retrosynthesis

**Support Vector Machines**

- Very good for extreme cases
- Support vectors brush up against the separating margin
- Considers only support vectors important
- Hyperplane - separates the classes in n-dimensional space
- Non linear data transformed to higher dimension
- Kernel - turns vectors to dot product in feature space
- Non-trivial to choose kernel
- **Applications** - medical imaging, air quality, medical classification, financial analysis, page ranking
**Medicinal Chemistry**
- Best Molecule
- Optimal Conditions

**Process Chemistry**
- Best Synthesis
- Many Reaction

---

**Merck: A Case Study**

- Process chemistry requires extremely exacting conditions in scale up
- Need quick access to a large quantity of multiple parameter permutations

---

**Medicinal Chemistry limited primarily by time and availability of materials**

Began in house 'kits' with predispensed catalyst and base to allow rapid evaluation of conditions

These kits enabled broader patterns to emerge from higher levels of data, leading to the Merck 'Catalyst Selection Guide'

Developed 'parallel-in-parallel' HTE - running various reaction conditions simultaneously with distinct nucleophile electrophile combinations

Bottlenecks lead to development of miniaturization platform - inspired by adjacent fields, unifying robotics in biotechnology with mass spec techniques

Achieved 1500 rxns in a day, with 0.02 mg per reaction driven by TPP Mosquito robot

---

**Acc. Chem. Res. 2017, 50, 2976**
Merck: NanoSAR

HTE enables rapid evaluation of potential compound leads
Greatest advances have come from a confluence of emerging technologies
to enable rapid evaluation and data collection on diverse compound sets

Merck developed NanoSAR for synthesis and affinity ranking using
UPLC-MS (Ultra-high-Performance Liquid Chromatograph/Mass Spec) and
ASMS affinity ranking (technique to detect compound bound to protein)

Workflow
1) Nanoliter robot dosing and reaction proceeds 20 hours
2) a. Confirm product via UPLC-MS b. info for up to 1536 rxns per run
3) Incubate protein with pool of molecules
4) a. Size exclusion column separates out unbound molecules b. HPLC-MS
   identifies complexes
5) Repeat ASMS assay at lower protein concentration
6) Scale up compounds and determine IC₅₀ by functional assay

Essential Controls

123 mg to run 3114 reactions,
evaluate 384 nucleophiles,
and assay material

Nature, 2018, in press
History of Pursuit

Matthew O'Neill
Cornells Group Meeting 1.06.2018

Why so difficult (taken from Gryzowski)

1. Exceptions to the norm abound - 'Black Swans'

2. Automation was not easy

3. Molecular context is everything!

4. Stereochemistry was hard to define (shown below in SMARTs)

5. Hard to define positions - cost associated with every possible move

6. Size of search and lack of intelligent algorithms

Paraphrasing Churchill's famous words after the Allies' first major victory over the Axis forces in Africa, it is not the end, it is not even the beginning of the end, but it is the end of the beginning for the computer-assisted synthesis planning. The machine is here to stay. - Gryzowski 2016
"These DNN were trained on essentially all reactions ever published in organic chemistry."

Synthetic decision trees utilize 3N-MCTS

- green - target
- red - readily available starting materials
- grey - low scoring routes

A bigger part is the supposition by organic synthetic chemists that they are the ones best qualified to design programs for organic synthesis. Not sol - Whitesides
Abby Doyle created first general machine learning algorithm for homogenous catalysis

Machine learning is exceptionally hard to amend to available data

Logistical constraints have largely precluded genuine multiparameter optimization in academia

Invokes the 'curse of dimensionality' - each additional parameter $n$ adds another dimension to the reactivity space

In typical methodology development, only a small fraction of this chemical space is thoroughly explored

Alternative - Use HTE to explore large amounts of chemical space
Can these larger data sets be used to teach algorithms how to predict yields when using unseen reagents?

Buchwald-Hartwig Amination

\[
\text{ArNH}_2 + \text{C=C-R} \xrightarrow{\text{Pd catalyst (10 mol%), additive (1 equiv)}} \text{NH-C=C-R}
\]

Aryl Halides (15) Additives (23) Bases (3) Pd Catalysts (4)

1536 well plates with Mosquite robot - UPLC to determine yield - 4608 rxns

Based on Glorious 'robustness screening'
Using additive approach to model embedded isoaxazole to simplify process

Create descriptors without invoking a mechanistic hypothesis, only look at structure
Internal consistency achieved using calculated DFT properties (B3LYP/6-31G*)

Workflow for Data Input

1) Submit molecular structure to Spartan GUI and specify reaction component
2) Calculation of molecular, vibrational, and atomic properties
3) Extraction of these features from the resulting text files
4) Generation of data table for fit - 120 total descriptors

Supervised ML Tested

$k$-nearest neighbors, support vector machines, and Bayes generalized linear models all no better than linear regression
Random forest model proved significantly better! 7.8% RMSE with $R^2=0.96$ - large amount of error attributed to experimental and analytical error
Better than any other method when trained on 5% of data vs. 70% of data for others

Examine Descriptors to Identify Most Important Qualities

Mechanistic insights from relative importance of descriptors
1) $^1$H NMR shift 2) LUMO energy 3) $^1$O1 charge 4) $^1$C5 charge

Based on Glorius 'robustness screening'
Additional Methods

Matthew O’Neill

Deoxyfluorination - Second Generation work from Doyle

- Fewer reactions - 640 with stock solutions and NMR quantification
- No mosquito robot :(
- Yields over 100% were kept to avoid bias - reveals experimental error as strong influence on subsequent prediction performance
- Reaction parameters not considered - time, temp, concentration, stoichiometry

Data input different - extract 7 features of molecules, and add categorical labels for a total of 23 different descriptors

Alcohol - *C1 charge, *C1A2, EN
Base - *N1
Sulfonyl Fluoride - *S1 charge, *F1 charge, *O1 charge

Yields for all 640 rxns determined by 19F NMR

Categorical Descriptors
- Alcohol - primary
- Alcohol - secondary
- Alcohol - tertiary
- Alcohol - cyclic
- Alcohol - 4-membered ring
- Alcohol - 5-membered ring
- Alcohol - 6-membered ring
- Alcohol - 7-membered ring
- Alcohol - benzylic
- Alcohol - allylic
- Alcohol - homobenzylic
- Alcohol - homoallylic
- Alcohol - alpha-carbonyl
- Alcohol - beta-carbonyl
- Alcohol - hemiacetal
- Alcohol - amino alcohol

Automated Closed-Loop Optimization Platform

Hepsin Inhibitor - clinical serine protease for carcinogenesis and metastasis via HGF signaling

Closed loop only takes 1-2 h to complete - traditional methods would take weeks

CyclOps platform for rapid synthesis vs. batch methods

BOUS - Best Objective Under-Sampled - algorithm predicts IC_{50}

Hepsin & uPA assays
Concentration data for assay

Merck QSAR challenge in 2012

Precomputed molecular descriptors for compounds and provided experimental biological activity for 15 targets; DNN model won the competition, without a single chemist on the team

J. Med. Chem. 2018, 61, 4335

My own feelings are that any machine that takes away from me the necessity to crank out 142 amide reactions by hand is welcome to it - Derek Lowe
NMR and X-Ray largely replaced degradation studies as preferred method for structure elucidation - is this bad? Led directly to the 'Golden Age of Natural Product Total Synthesis' - this beautiful field would have been precluded by technological constraints were it not for the advent of NMR and other enabling analytical techniques.

“The age of automation” thus appears to hold the potential to advance organic synthesis in a revolutionary way” - Nuno Maulide

What will this revolution look like?

Innovation for this talk

“This evolution – guided by experts in artificial intelligence more than by experts in synthesis – raises a very important point for students: viz., for the future, will it be more important to understand AI than to be able to recall all the methods for introducing chiral centers (or other transformation of choice) from human memory? How should one balance computer skills and empirical synthetic skills?” - Whitesides Isr. J. Chem. 2018, 58, 142.

Coming Full Circle

What defines a practitioner of synthesis?

"I did not become a “synthetic organic chemist,” but almost all the research that my colleagues and I have done (and do) involves organic synthesis." - Whitesides

Satire Article about Martin Burke MD PhD Synthesis Machine - Science, 2015, 347, 1221

Urbana, IL

Ugly scenes today marred the unveiling of what may become a landmark paper, as an angry mob of organic synthesis researchers invaded the chemistry department at the University of Illinois at Urbana–Champaign before seizing and ultimately destroying a so-called "synthesis machine."

An article in the journal Science, describing the development of what is in effect a cyborg post-doc, prompted an initially peaceful protest outside the chemistry department under placards carrying the slogans KEEP NATURAL PRODUCT SYNTHESIS NATURAL, SUZUKI COUPLINGS ARE CHEATING and GIVE ME C–H ACTIVATION OR GIVE ME DEATH. However, witnesses described a marked increase in tension after the arrival of a counter-demonstration of inorganic chemists, who taunted their organic counterparts with highly charged epithets including "pot-boiler" and "column monkey".

An anonymous demonstrator later told C&EN Onion: "It all kicked off when the fucking stamp collectors showed up. The was always an undercurrent of anger, but that was when it boiled over and you became keenly aware just how many people had brought BuLi with them."

Asked to explain the motives of the inorganic counter-demonstrators, a hooded organometallic researcher said, "We're just here looking for trouble. I've got no dog in this fight, unless you're gonna tell me that thing's got an onboard SQUID magnetometer."

Anger having now reached fever pitch, a large group stormed the building, making directly for the lab housing the controversial machine. Minutes later, the helpless automaton was flung from a second floor window, landing amongst cheering protesters and breaking, ironically enough, into a number of fragments. Amid frantic shouts that the machine may have developed the capability to heal itself, clamp-stand-wielding synthetic chemists smashed what little remained. To their credit, many of them first donned appropriate personal protective equipment.

John Wiseman, a technician present during the break-in, remained sanguine as he detailed the damage to the lab. "The automated synthesis platform was what they came for, of course, but someone also found time to steal a bunch of NMR tubes and a fresh batch of DMP. You know what these people are like."

Wiseman also claimed that clashes involving armed factions of researchers were not without precedent: "You'd be surprised. There are a lot of radical chemists out there."

Taken from http://cenonion.blogspot.com/2015/03/rampaging-synthetic-chemists-smash.html
Appendix

Closed Look Hepsin Inhibitors

**Depiction of Flow Process**

Commercially available inhibitors - identify 7 as lead molecule

---

**Predicting Reaction Performance in C-N couplings using ML**

**Examples of calculated descriptors**

Additive Descriptors (n = 19)
- EHOMO, ELUMO, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular
- Aryl Halide Descriptors (n = 27)
- EHOMO, ELUMO, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular
- Base Descriptors (n = 10)
- EHOMO, ELUMO, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular
- Weight, Ovality, Surface Area, *N1 Electrostatic Charge
- Ligand Descriptors (n = 64)

---

<table>
<thead>
<tr>
<th>Compound Number</th>
<th>Structure</th>
<th>Hepsin IC₅₀ (µM)</th>
<th>8PA IC₅₀ (µM)</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>&gt;10</td>
<td>ND</td>
<td>2.43</td>
<td>0.676</td>
<td>4.46</td>
<td>6.55</td>
<td>0.272</td>
<td>6.19</td>
<td>3.35</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>&gt;10</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.029</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.383</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>4.02</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.005</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>1.13</td>
<td>&gt;10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>