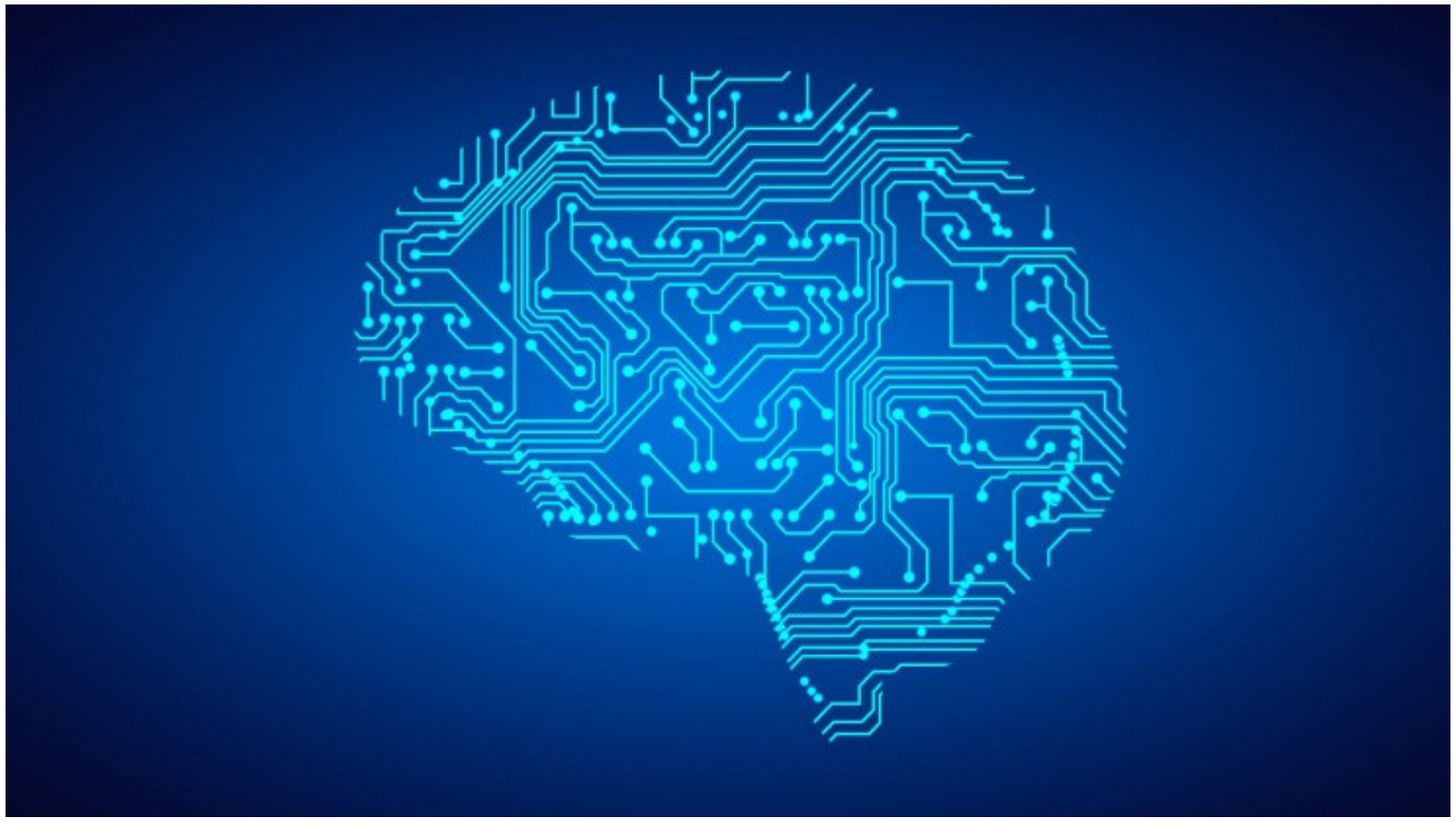
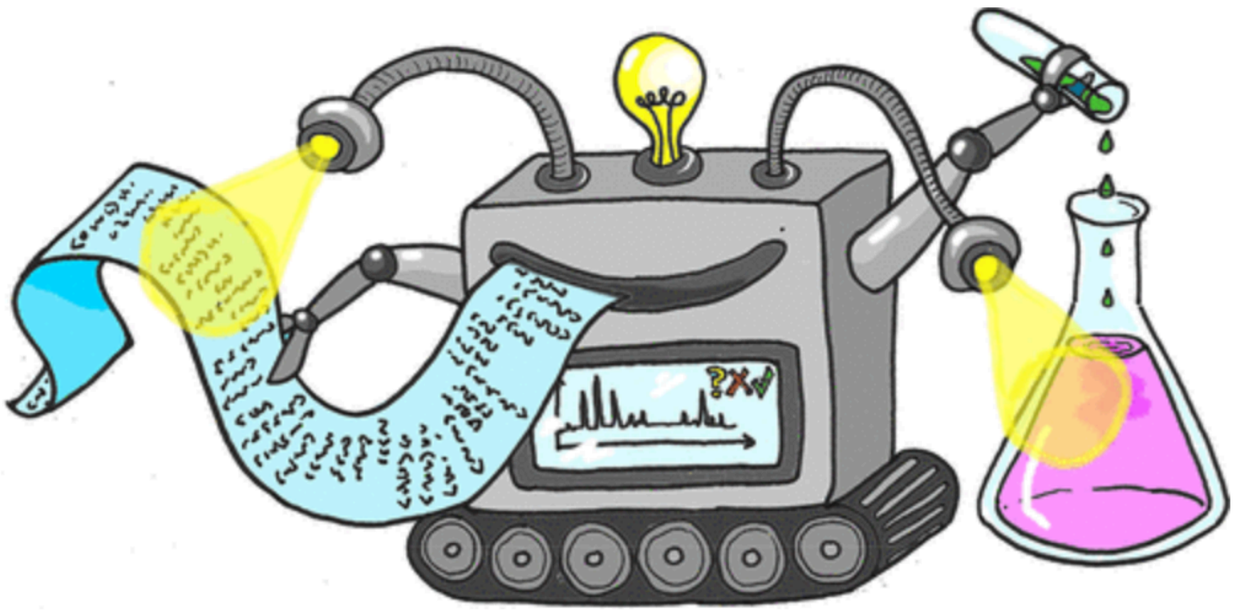
Machine learning



Jacob Kautzky
Group Meeting
12-5-18

Machine learning outline



What is machine learning?

Common machine learning algorithms

Machine learning applied to chemistry

- Retrosynthesis
- Reaction discovery
- Reaction optimization
- Drug optimization
- Materials chemistry

What is machine learning?

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with the experience E.

- Branch of artificial intelligence
- Design systems that learn and improve as data is gathered
- The system is utilized to refine a model that in turn can predict outcomes

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Data gathering

Collect the data yourself

need to gather a sufficient quantity of data

- High throughput
- Automation

Get the data from other sources

In the past (and to lesser extents today), it was difficult to convert data from published papers and patents into data that could be utilized for machine learning

Datasets can have untrustworthy data and a way needs to be devised to prevent them from overly influencing the conclusions

Published papers generally omit data that does not work, which can cause problems in trying to build complete datasets

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Supervised vs unsupervised learning

Supervised learning — working with a set of labeled training data where each piece of data has an input and output object

- Decision trees
- Logistic regression
- Support vector machine
- Random forest
- Naive Bayes
- Relevance vector machine
- Linear regression
- Artificial neural networks
- k nearest neighbor

Unsupervised learning (data mining) — the algorithm finds hidden patterns in a load of data

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Algorithm types: linear and logistic

Most basic forms for regression and classification respectively

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Algorithm types: linear and logistic

Most basic forms for regression and classification respectively

Linear - fitting a polynomial curve to data

- Easily interpretable results
- Fast to train a machine
- Low average prediction accuracy
- Easily skewed by outliers

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

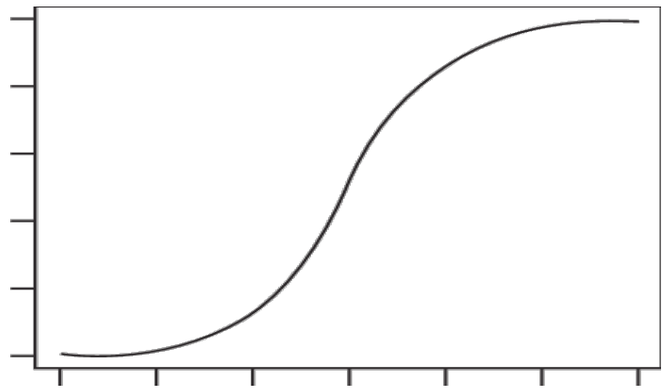
Algorithm types: linear and logistic

Most basic forms for regression and classification respectively

Linear - fitting a polynomial curve to data

- Easily interpretable results
- Fast to train a machine
- Low average prediction accuracy
- Easily skewed by outliers

Logistic - measures the relationship between the categorical dependent variable and one or more independent variables by using a logistic function



Binary logistic curve

Similar to a linear model, however the outputs are restricted between 0 and 1 and there is a different conditional basis set (Bernoulli vs Gaussian)

- Similar properties to linear, but slighly more difficult for the user to interpret the results

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

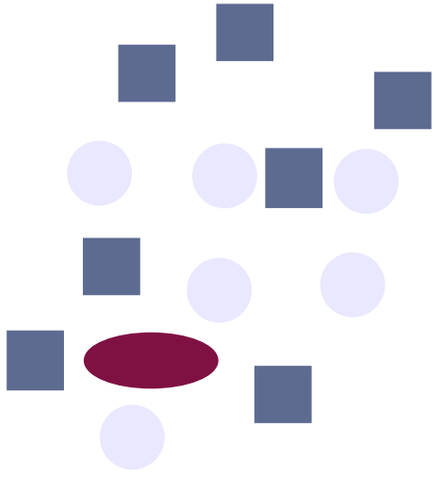
Algorithm types: k nearest neighbor

"Lazy learning" method that is dependent on the local structure of the data

For classification, the object is classified by a majority vote of its k closest neighbors

For regression, the value is the average of its k closest neighbors

if k = 5, then the unclassified maroon oval would be a slate blue square, but if k = 8 it would instead be a lilac circle



- Easily interpreted results
- Low average predictive accuracy
- Poor ability to separate signal from noise

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Algorithm types: naive bayes classifier

- Family of probabilistic classifiers
- Assume independence between the features
- Apply Bayes' theorem

Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Note: P(A I B) is the probability of A given B

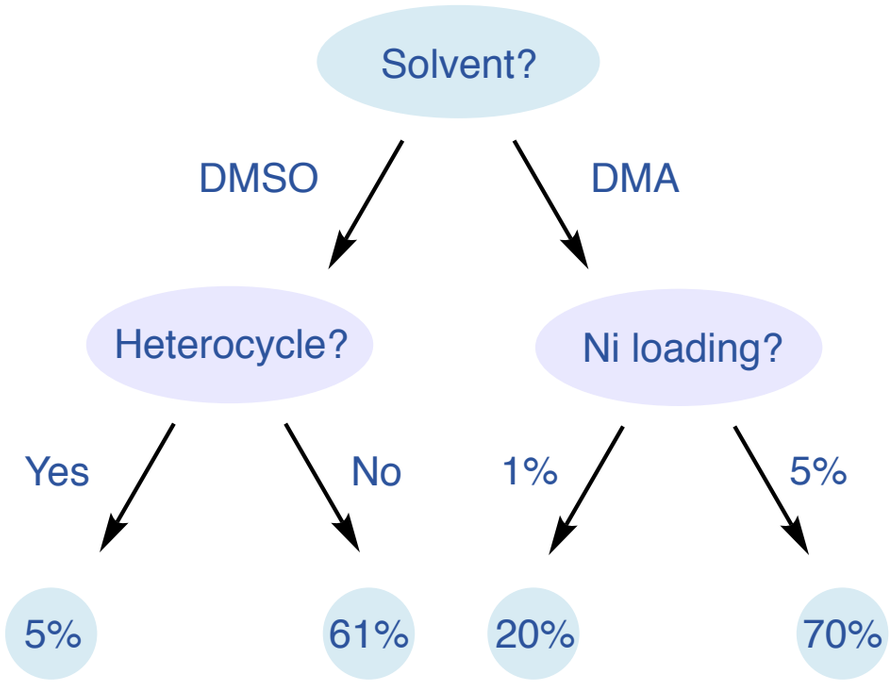
Despite their apparent oversimplification, naive Bayes classifiers have worked quite well in several real world applications

The advantage of a Bayes model is that it requires a small amount of training data necessary to attain the parameters required for classification, however it suffers from slightly lower average predictive accuracy in several instances

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Algorithm types: decision tree

A form of sorting whereby a tree is split into branches by the imput variables



Advantages

Fast training speed

Fast prediction speed

Results somewhat interpretable by user

Can be used for both classification and regression

Disadvantages

Generally low prediction accuracy

Frequently overfits the data

Performs poorly with few observations

Poor ability to deal with outliers and separate signal from noise

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Algorithm types: random forests

Similar to a decision tree

Rather than making just one tree, several trees are produced, with each of being randomly selected to take into account only certain variables

Select the mode classification or mean prediction from all of the individual trees

- Decreases the tendency to overfit the data
- Good at separating signals from noise
- The reasoning behind the algorithm is more difficult for someone to understand
- Can be skewed toward certain descriptors if there are more variables associated with a given descriptor

Can be avoided by using the cforest algorithm which can avoid descriptor selection bias

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

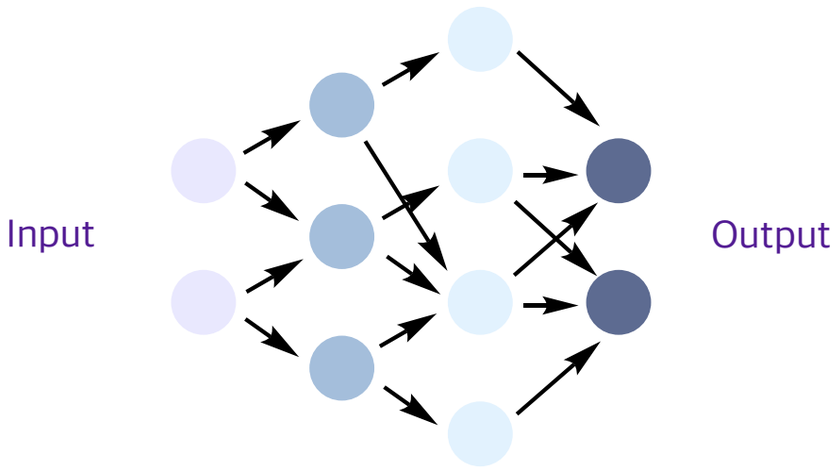
Algorithm types: artificial neural network

Computing system inspired by biological neural networks

A framework for many machine learning algorithms to work together and process coplex data imputs

Based on a collection of connected nodes called artificial neurons

An artificial neuron can transmit a signal to other neurons attached to it by an edge that in turn can signal other neurons



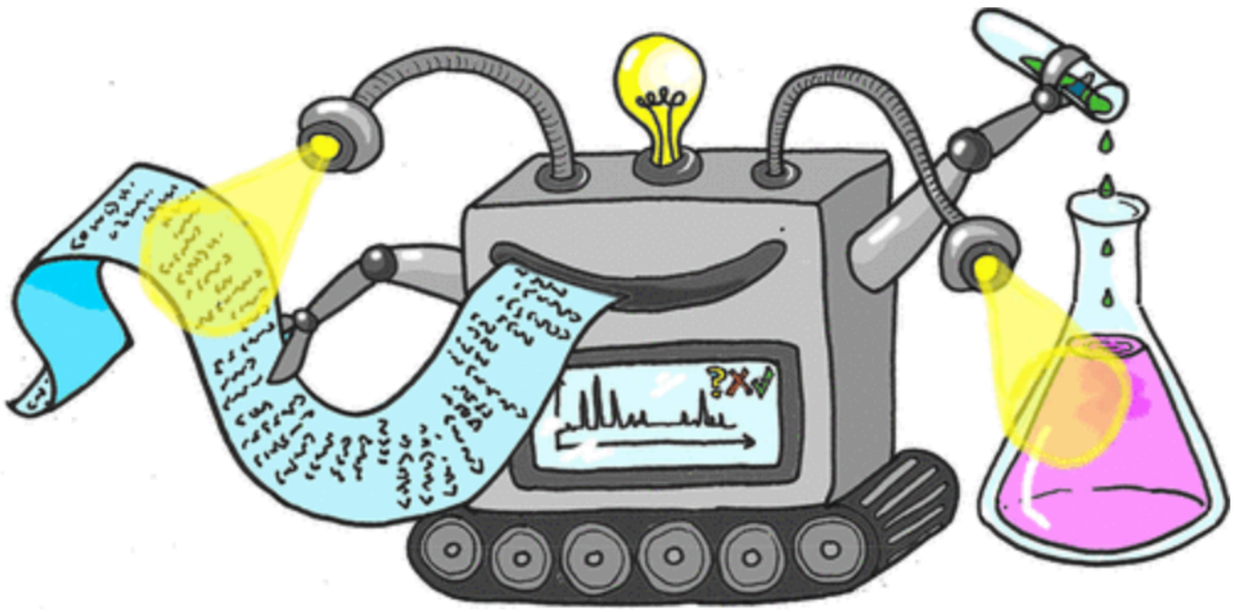
Examples of this include learning to identify images that contain cats by being shown images that have been manually labeled cat or no cat

This has been used by groups in the past to convert chemical structures in patents and old papers into SMILES and other readable formats for machine learning

- Challenging to understand what the algorithm is doing
- Deals with irrelevant features well
- Slow training speed
- High average predictive accuracy

Bell, J. *Machine Learning: Hands-On for Developers and Technical Professionals; John Wiley & Sons, Inc.: Indianopolis, IN, 2014.*

Machine learning outline



What is machine learning?

Common machine learning algorithms

Machine learning applied to chemistry

- Retrosynthesis
- Reaction discovery
- Reaction optimization
- Drug optimization
- Materials chemistry

Machine learning

Retrosynthesis

Reaction optimization

- Identifying ideal condtions
- Predicting yields

Medicinal Chemistry

- Structure generation
- Binding prediction
- Estimating ADMET properties

Chemistry

Reaction discovery

Materials chemistry

- Identifying ideal alloys
- Synthesis/ Crystalization

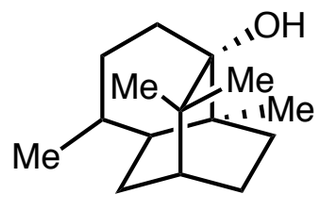
History of machine learning and synthesis planning

Corey

Wipke

1969: Organic Chemical Simulation of Synthesis (OCSS)

First computer assisted retrosynthetic analysis

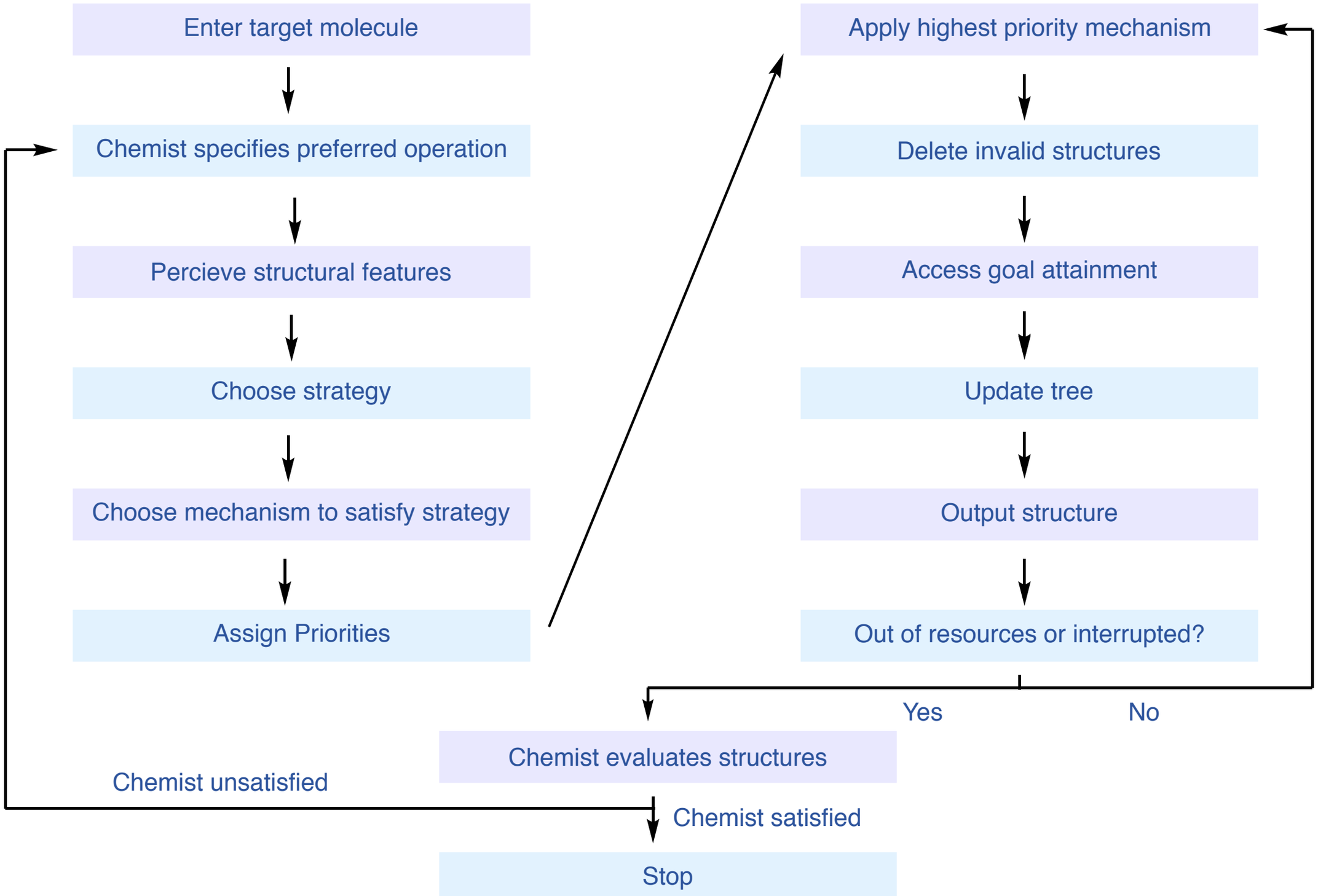


Provided a list of heuristic conditions to guide the choice of synthetic disconnections

Short lived and eventually split in two

Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178.

Organic Chemical Simulation of Synthesis (OCSS)



Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178.

History of machine learning and synthesis planning

1976: Wipke's SECS

1969: OCSS

1972: Corey's LHASA

- Built upon the LHASA approach by expanding the database
- Received backing from Swiss/German pharma companies, but eventually disconnected

Provided a set of six strategies for retrosynthetic analysis

The chemist chooses which strategies to be tried

"LHASA was not designed to invent chemistry that has never been performed in a laboratoy" - Corey

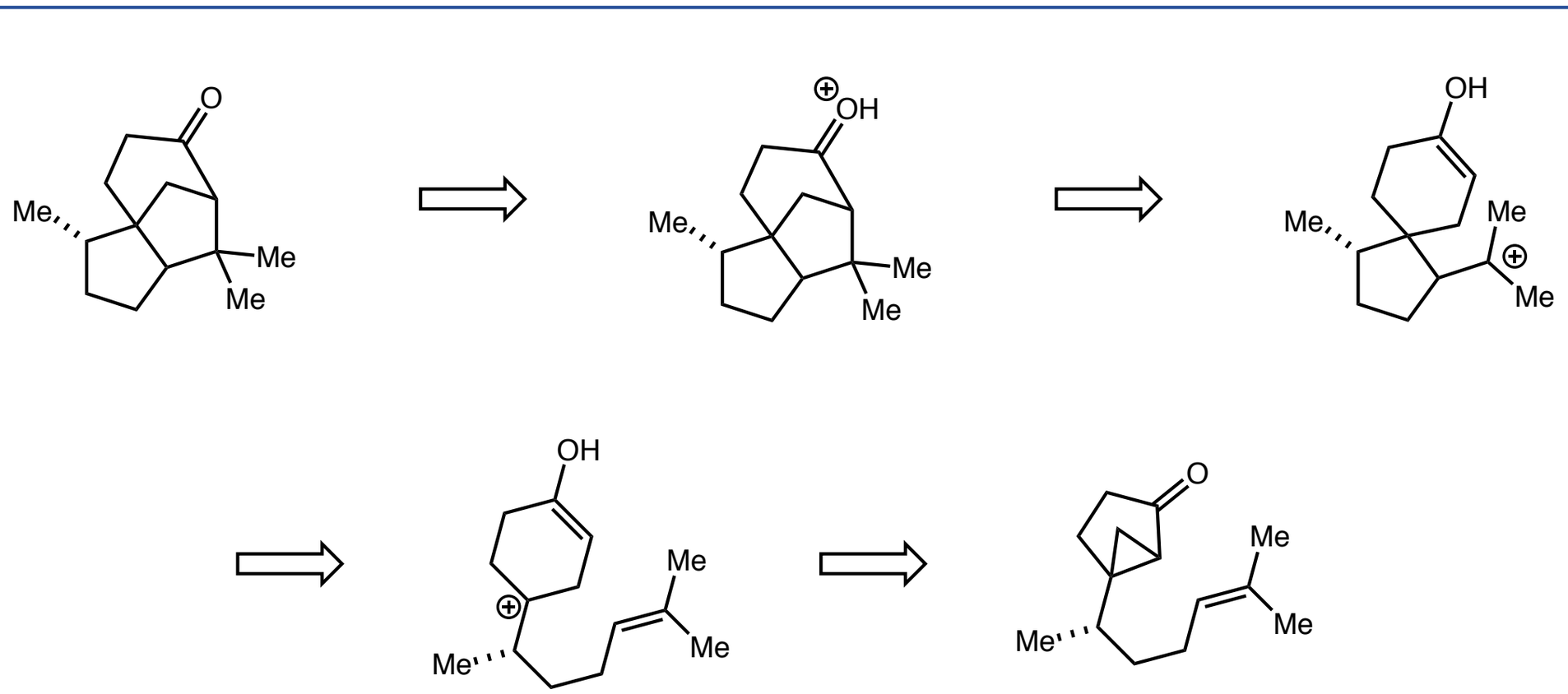
Continuously updated through 2010

Grzybowski, B. A. et al *Angew. Cem. Int. Ed.* **2016**, *55*, 5904.

Six principles of Logic and Heuristics Applied to Synthetic Analysis (LHASA)

Transformation-based strategy — Identification of a powerful simplifying transformation (ex. Diels- Alder, Aldol cyclization, etc.)

Mechanistic transforms — The target is converted to a reactive intermediate from which other intermediates of synthetic value can be generated

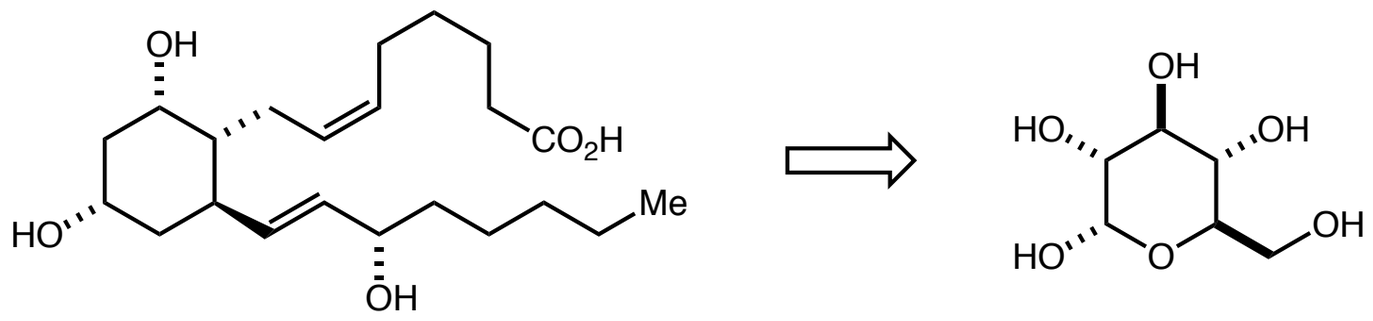


Sample mechanistic transformation

Corey, E. J.; Long, A. K.; Rubenstein, S. D.. *Science* **1985**, *228*, 408.

Six principles of Logic and Heuristics Applied to Synthetic Analysis (LHASA)

Structure-goal — Identification of a potential starting material, building block, retron-containing subunit or initiating chiral element



Topological strategies — Identification of one or more bonds that could lead to major simplifications

Stereochemical strategies — Stereoselective reactions, or steric based arguments used to reduce stereocomplexity

Functional group-oriented strategies — Functional group interconversions and determining logical disconnections based off of functional group arrangement(s)

Corey, E. J.; Long, A. K.; Rubenstein, S. D.. *Science* **1985**, *228*, 408.

History of machine learning and synthesis planning

1976: Wipke's SECS

Tried converting compounds into theoroms that needed to be proved

1969: OCSS

1989: Johnson's SYNLMA

Combinitorial explosion issues

1972: Corey's LHASA

1989: Hendrickson's SYNGEN

1977: Gelernter's SYNCHEM

- Self-guided
- Too "short-termed"

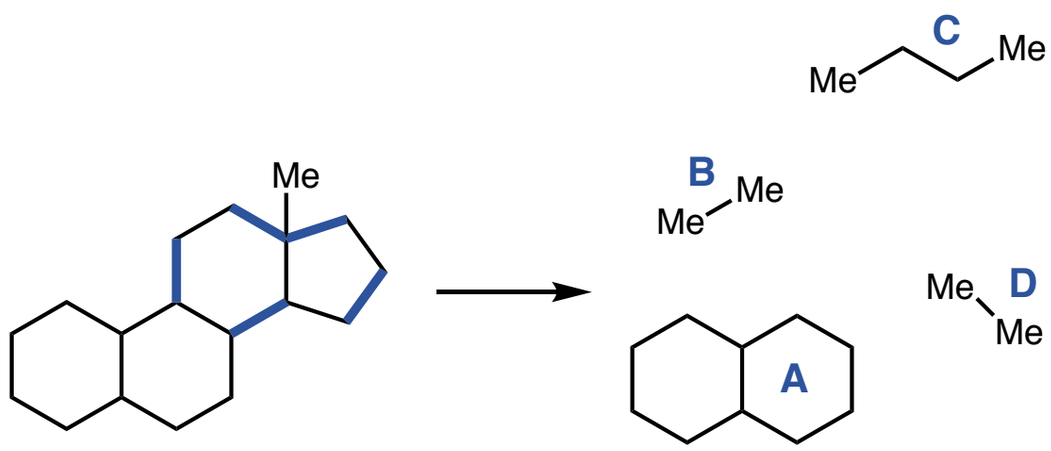
Grzybowski, B. A. et al *Angew. Cem. Int. Ed.* **2016**, *55*, 5904.

Hendrickson's SYNGEN

Focus on skeletal construction

Bondset — a set of bonds λ which need to be constructed

The number of ways to break apart a skeleton is equal to the number of possible bondsets



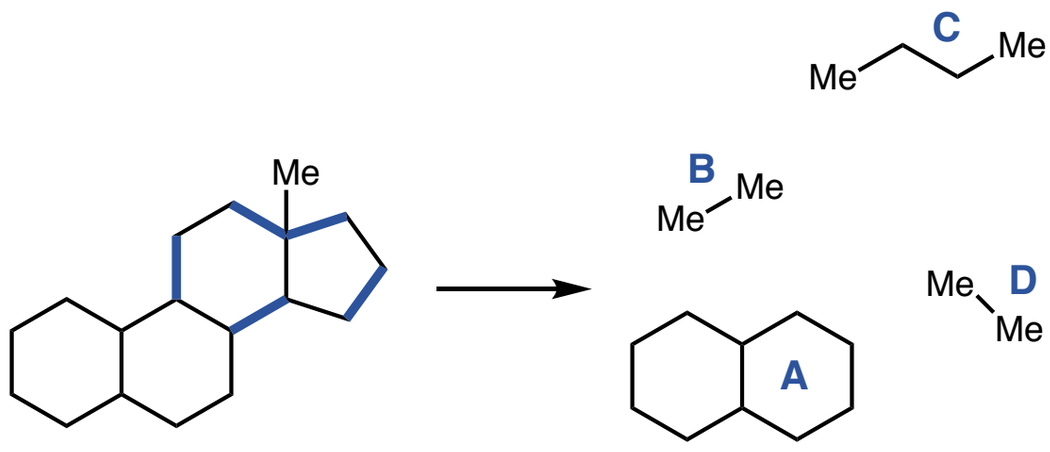
b! / (*b*-λ)! possible bond sets for *b* bonds

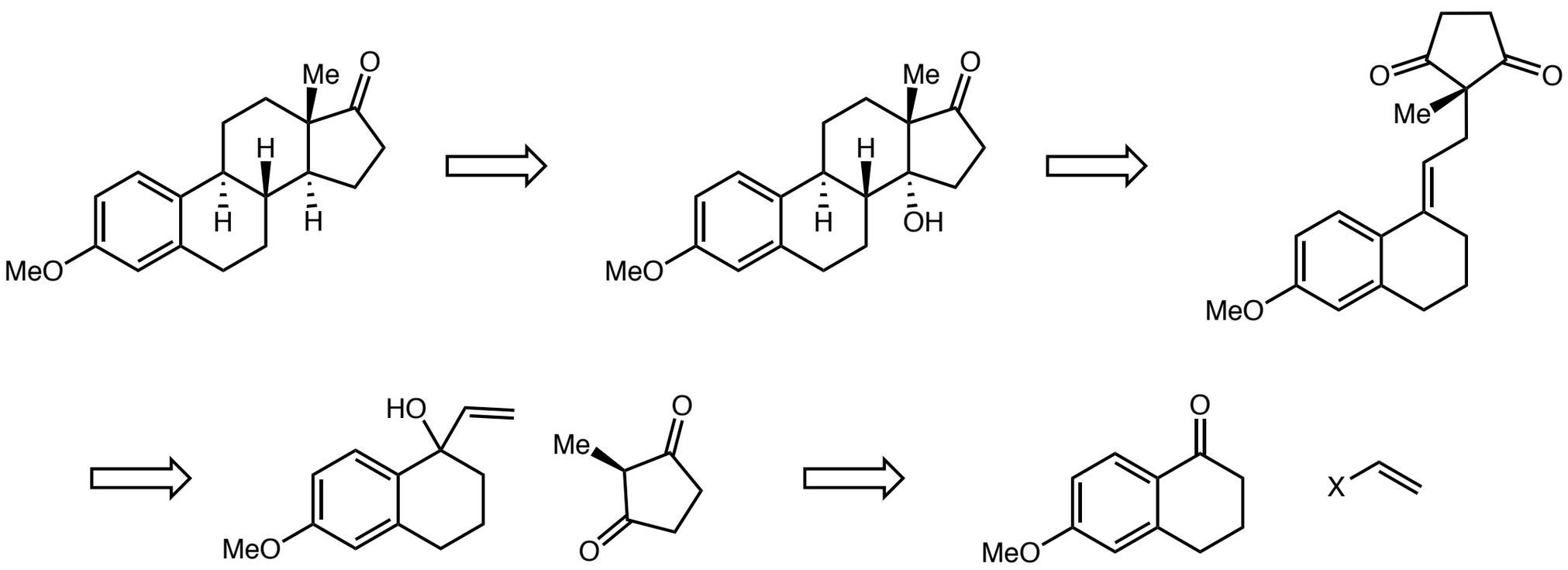
Hendrickson, J. B. *Angew. Chem. Int. Ed.* **1990**, *29*, 1286.

Hendrickson's SYNGEN

Focus on skeletal construction

Using an algorithm, SYNGEN strives for convergent assemblies using starting materials in its database





Hendrickson, J. B. *Angew. Chem. Int. Ed.* **1990**, *29*, 1286.

History of machine learning and synthesis planning

1969: OCSS

1972: Corey's LHASA

1976: Wipke's SECS

1977: Gelernter's SYNCHEM

1989: Johnson's SYNLMA

1989: Hendrickson's SYNGEN

1990: Hanessian's CHIRON - recognized sterochemistry and suggested methods to access it from a chiral pool

1993: Ugi's IGOR

- Bonds represented as bond-electron matrices
- Reactions treated as R-matrices attained by the subtraction of substrate and product matrices
- Able to identify novel pericyclic transformations
- Computationally intensive

1995: Gasteiger's WODCA - bidirectional synthetic planning assistor that focuses on the fundamental properties of bonds to suggest synthetic disconnections

Grzybowski, B. A. et al *Angew. Cem. Int. Ed.* **2016**, *55*, 5904.

History of machine learning and synthesis planning

1969: OCSS

1972: Corey's LHASA

1976: Wipke's SECS

1977: Gelernter's SYNCHEM

1989: Johnson's SYNLMA

1989: Hendrickson's SYNGEN

1990: Hanessian's CHIRON - recognized sterochemistry and suggested methods to access it from a chiral pool

1993: Ugi's IGOR

1995: Gasteiger's WODCA - bidirectional synthetic planning assistor that focuses on the fundamental properties of bonds to suggest synthetic disconnections

1998: Barone and Chanon's SESAM - identifies synthons based on skeletal overlaps, but ignores functional groups

2009: ARChem Route Designer - Machine extracted ~100,000 rules from literature examples that are utilized to design synthetic routes

Grzybowski, B. A. et al *Angew. Cem. Int. Ed.* **2016**, *55*, 5904.

Synthesis Planning with MCTS-3N

The purpose of a Monte Carlo tree search is to given a game state, choose the most promising next move

Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Synthesis Planning with MCTS-3N

The purpose of a Monte Carlo tree search is to given a game state, choose the most promising next move

MCTS Steps

- Selection
- Expansion
- Simulation
- Backpropogate

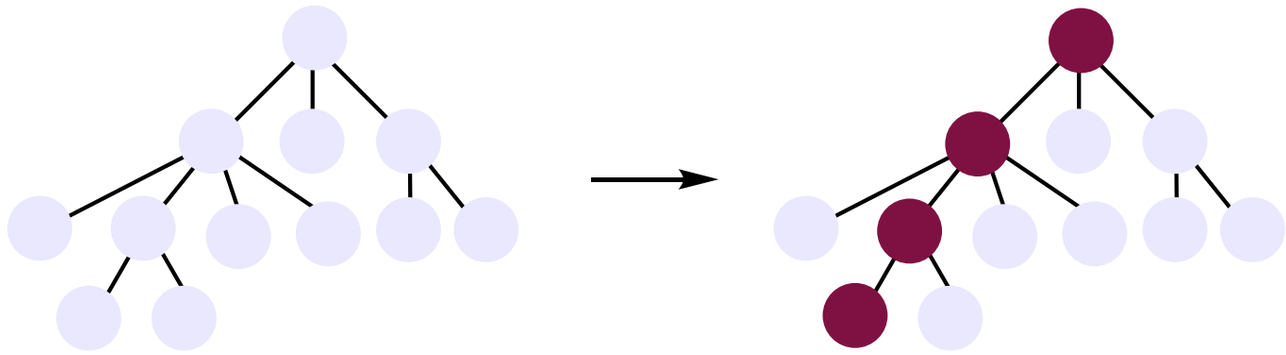
Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Synthesis Planning with MCTS-3N

The purpose of a Monte Carlo tree search is to given a game state, choose the most promising next move

MCTS Steps

- Selection
- Expansion
- Simulation
- Backpropogate



Select a route toward the most promising move

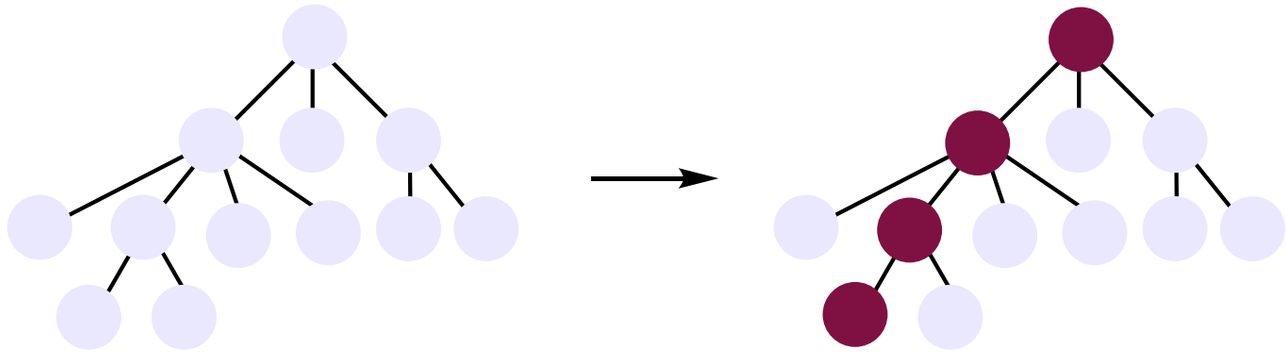
Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Synthesis Planning with MCTS-3N

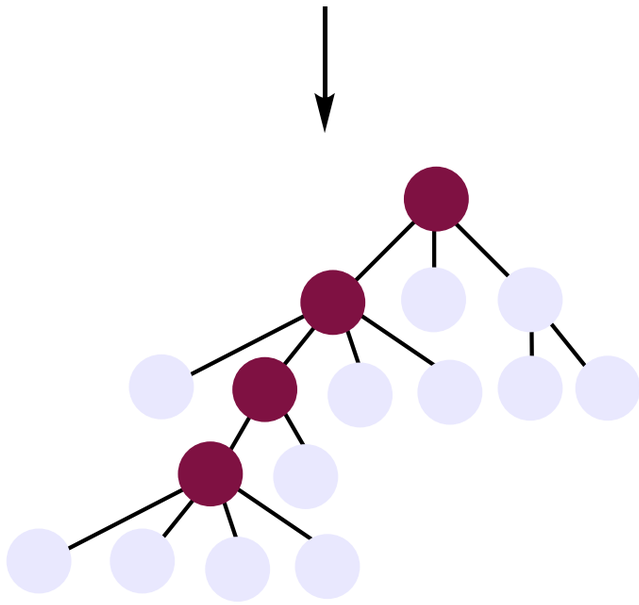
The purpose of a Monte Carlo tree search is to given a game state, choose the most promising next move

MCTS Steps

- Selection
- Expansion
- Simulation
- Backpropogate



Select a route toward the most promising move



Expand the tree by giving it further options off of its current leaf

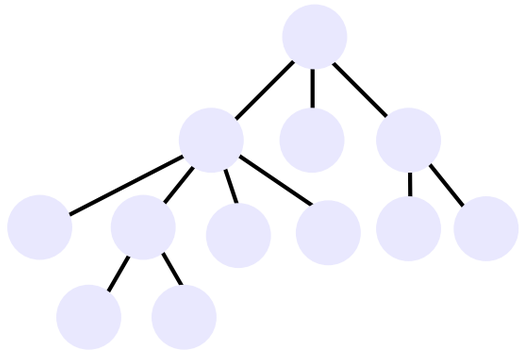
Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Synthesis Planning with MCTS-3N

The purpose of a Monte Carlo tree search is to given a game state, choose the most promising next move

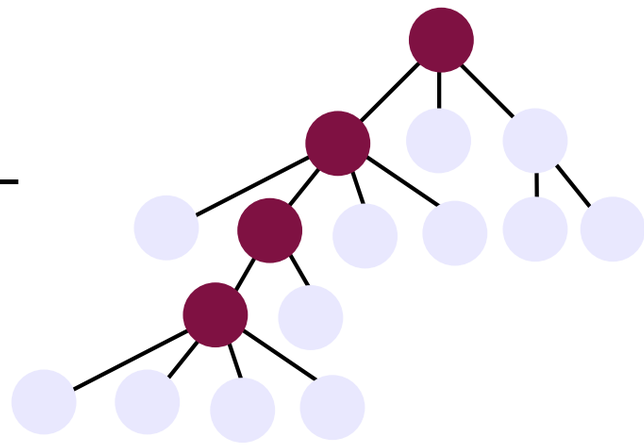
MCTS Steps

- Selection
- Expansion
- Simulation
- Backpropogate



Select a route toward the most promising move

Transmit the data back to the initial graph



Simulate each of the new leaves

Expand the tree by giving it further options off of its current leaf

This process is completed recursively until a route is found

Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Synthesis Planning with MCTS-3N

Merged a Monte Carlo tree search (MCTS) with three separate neural networks to devise retrosyntheses

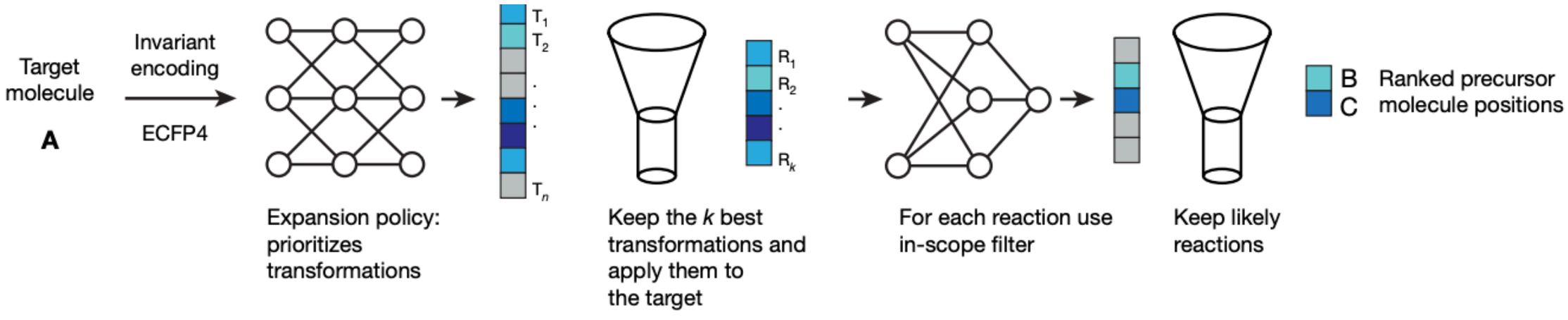
- Utilized all reactions on Reaxy's prior to 2015 to generate a training set
- Extracted 301,671 rules from this dataset and utilized them to train neural networks

  Neural networks learn the context in which reactions can occur (functional group tolerance)
- Developed three neural networks

  1. Expansion policy - guides the search in promising directions by proposing a restricted number of transformations

  2. Examine the feasibility of the proposed transformation

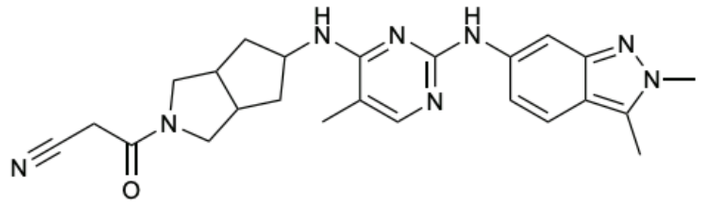
  3. Estimates the position value

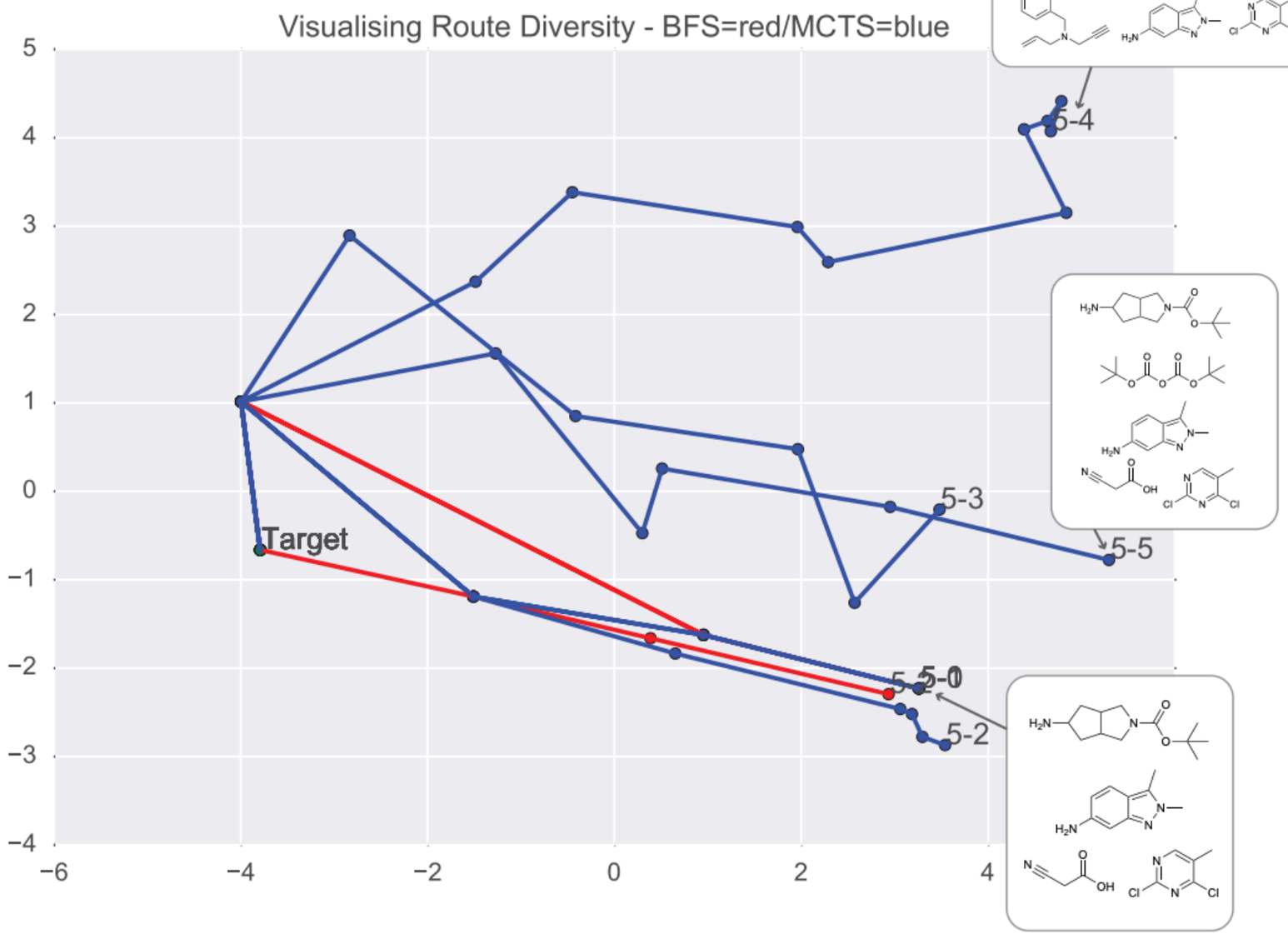


Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Synthesis Planning with MCTS-3N

Merged a Monte Carlo tree search (MCTS) with three separate neural networks to devise retrosyntheses





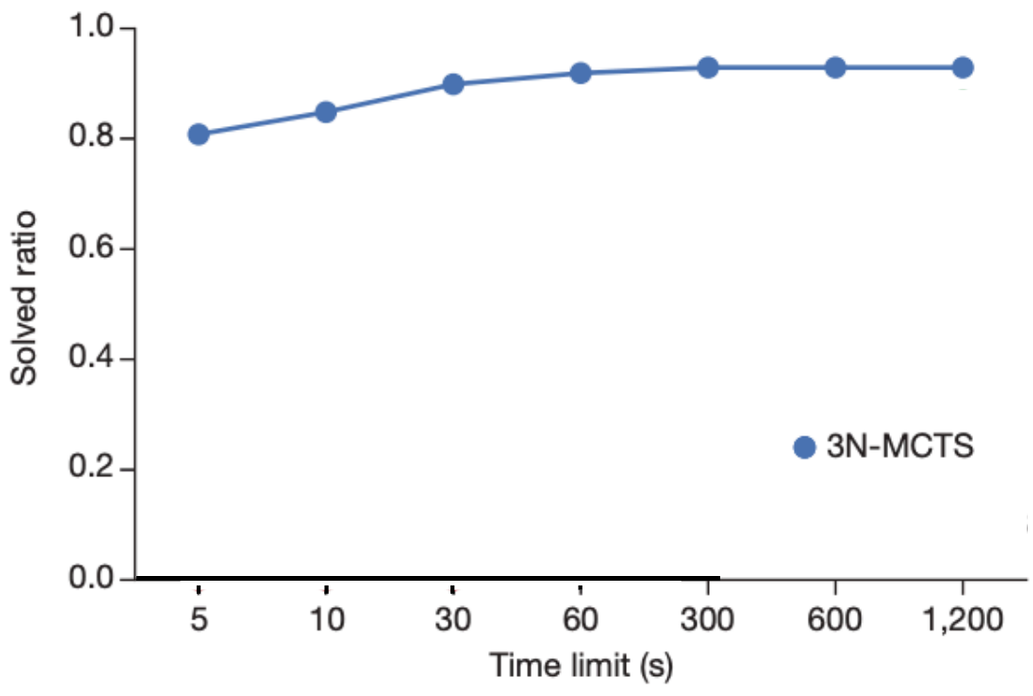
Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Synthesis Planning with MCTS-3N

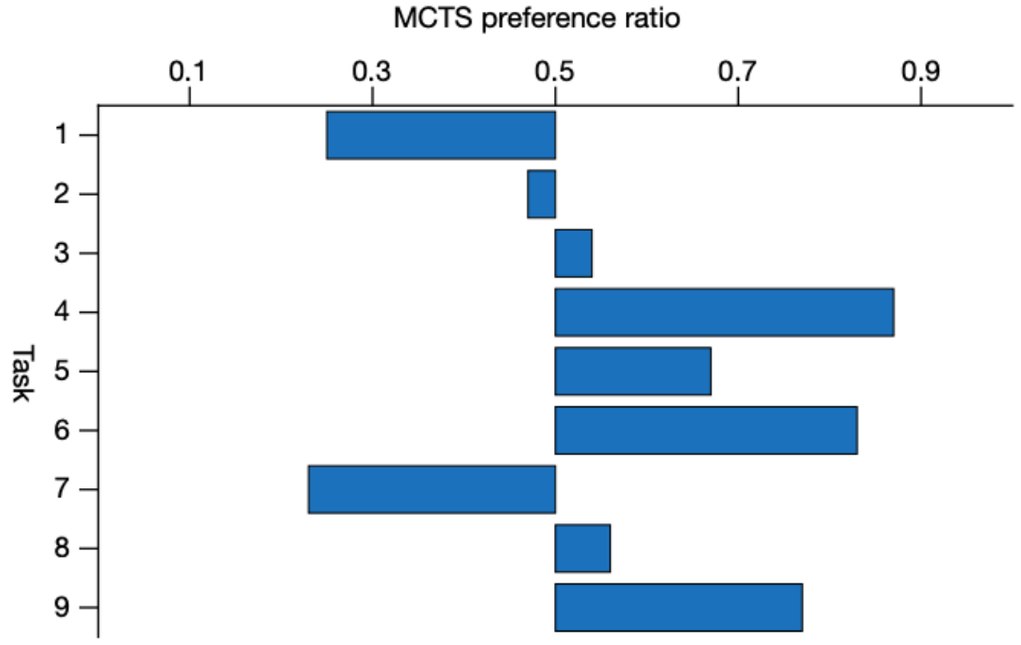
Evaluating Performance

- The neural network reaction checker has a false positive rate of 1.5% and false negative of 14%
- The neural network predicted the correct solution 31% of the time and had the correct solution in the top 50 73% of the time

The program was able to solve 80% of 500 diverse molecules in under 5 seconds



Double blind test asking chemists to pick between similarly long literature and MCTS syntheses

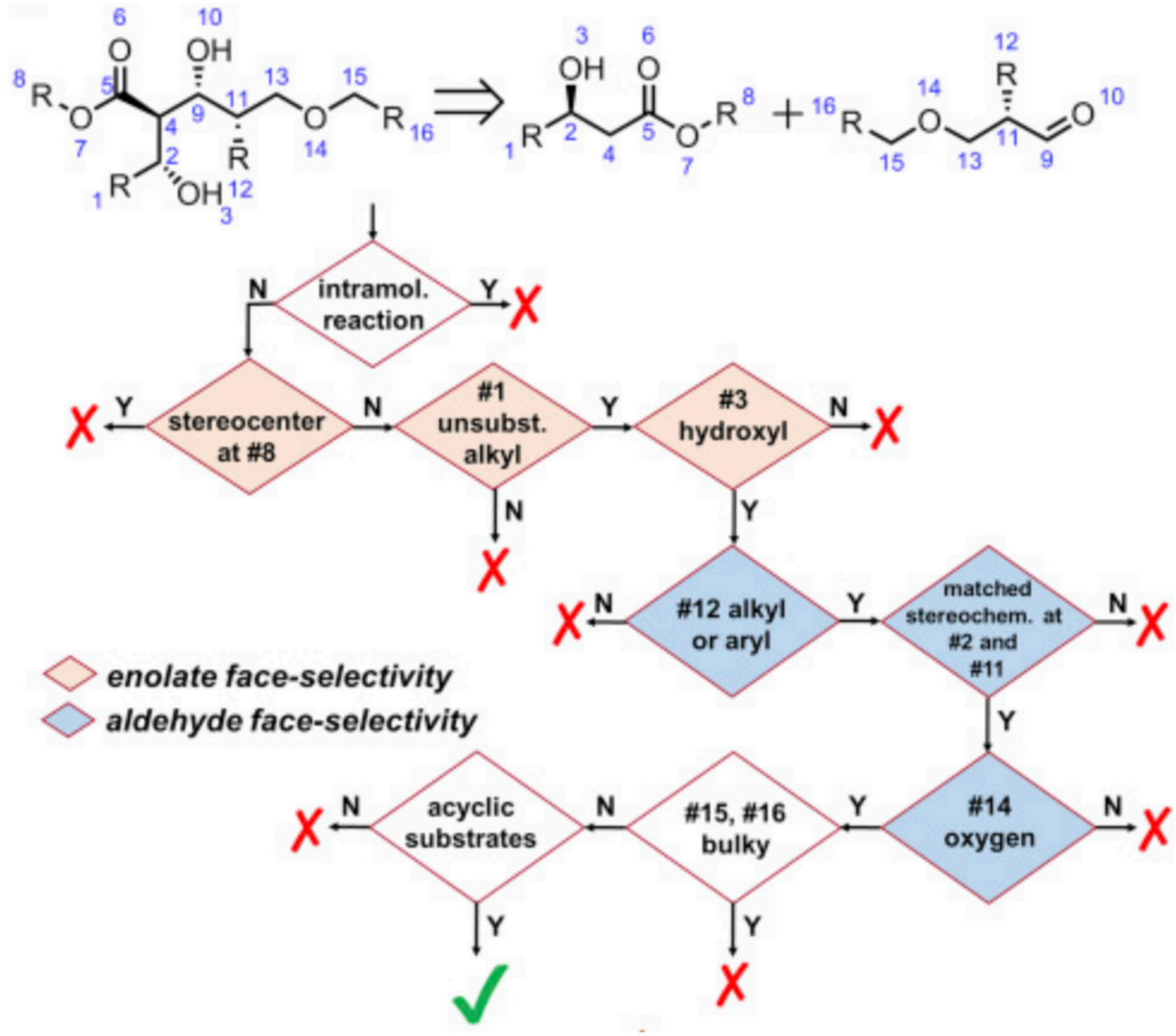


Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

Chematica

Unites network theory, modern high-power computing, artificial intelligence, and expert chemical knowledge to rapidly design synthetic pathways

Thousands of reactions hand coded by experts

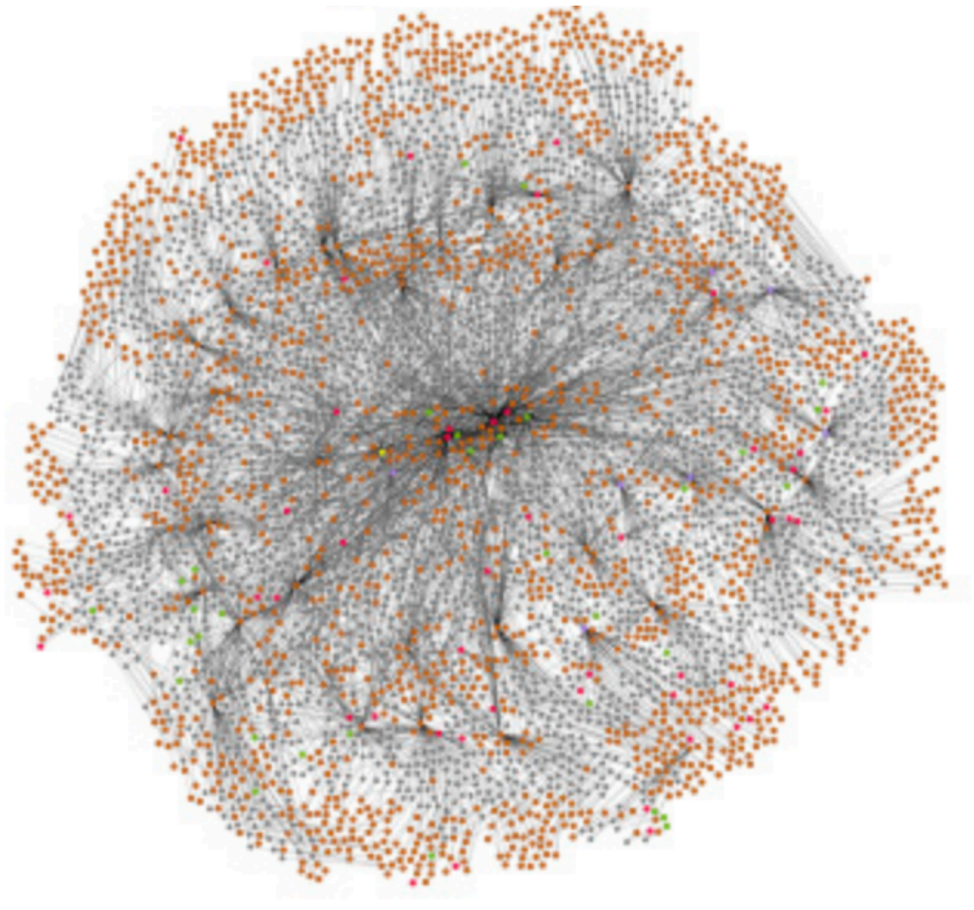


Grzybowski, B. A. et al. *Chem* **2018**, *4*, 522.

Chematica

Unites network theory, modern high-power computing, artificial intelligence, and expert chemical knowledge to rapidly design synthetic pathways

Thousands of reactions hand coded by experts



Simplified graphic example of the synthetic possibilities

Pruned down options by removing branches with unlikely structural motifs, that proceed through strained intermediates, etc.

A scoring algorithm evaluates the resultant substrates produced and the reactions required to make the set

Grzybowski, B. A. et al. *Chem* **2018**, *4*, 522.

Chematica

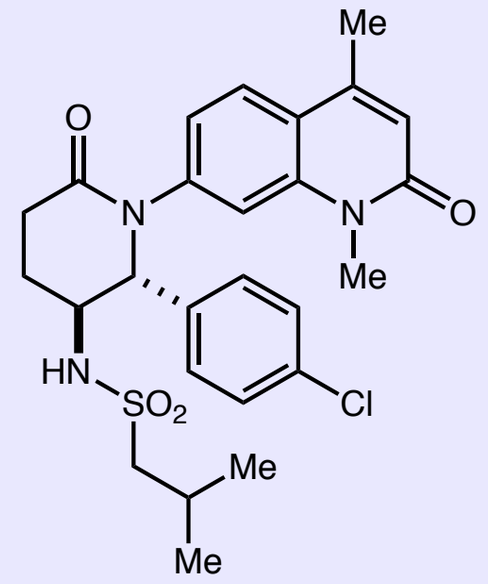
Unites network theory, modern high-power computing, artificial intelligence, and expert chemical knowledge to rapidly design synthetic pathways

Thousands of reactions hand coded by experts

The software was tested on several compounds of commercial interest to MilliporeSigma that currently had troublesome syntheses

The software designed routes that were then executed without changes except for straightforward adjustments to the reaction conditions (e.g. temperature, solvent, etc.)

In every instance improvements were made (shorter routes, fewer chromatographic steps, higher yields, more reproducible)



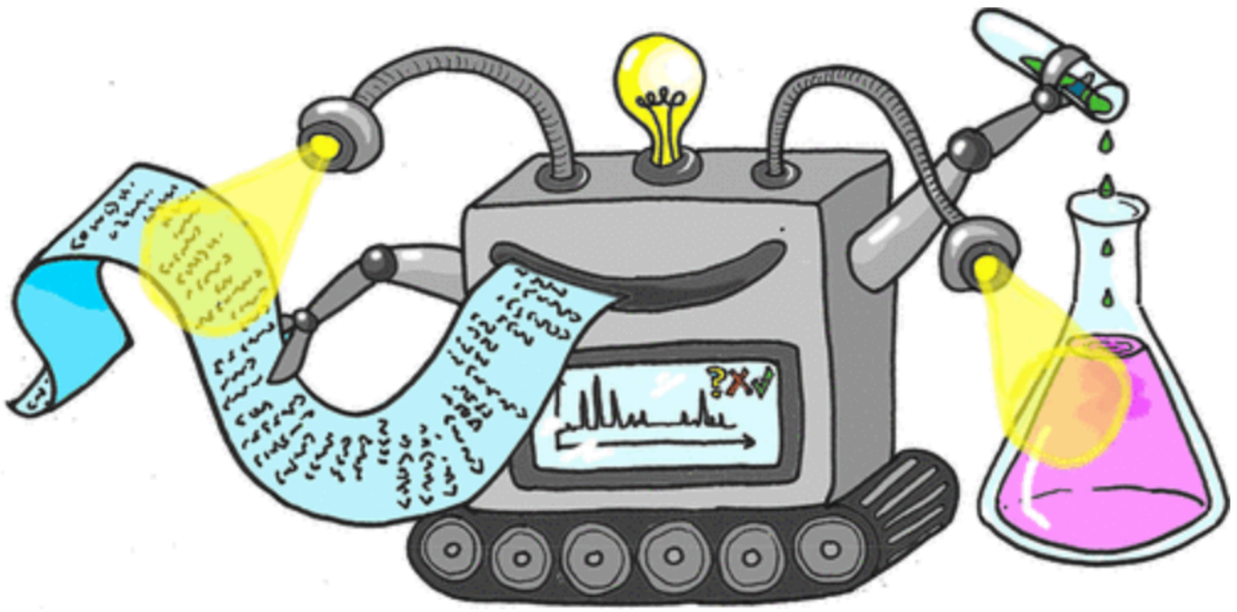
Improved the yield 8-fold relative to literature

Decreased the number of FCC separations

Chematica was also able to design a pathway that broke a patented route to a compound while more than doubling the yield!

Grzybowski, B. A. et al. *Chem* **2018**, *4*, 522.

Machine learning outline



What is machine learning?

Common machine learning algorithms

Machine learning applied to chemistry

Retrosynthesis

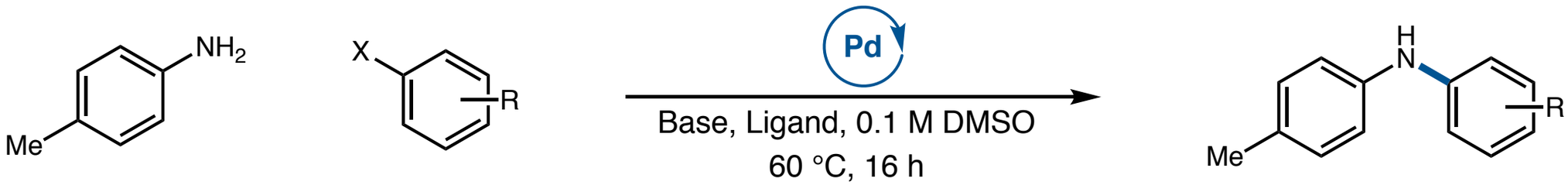
Reaction optimization

Reaction discovery

Drug optimization

Materials chemistry

Predicting reaction performance



Why are isoxazoles challenging in the Buchwald Hartwig reaction?

4608 reactions

15 aryl halides

23 isoxazole additives

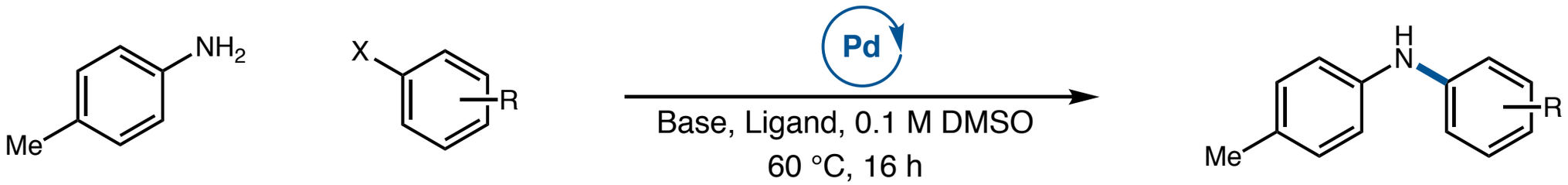
4 Pd catalysts

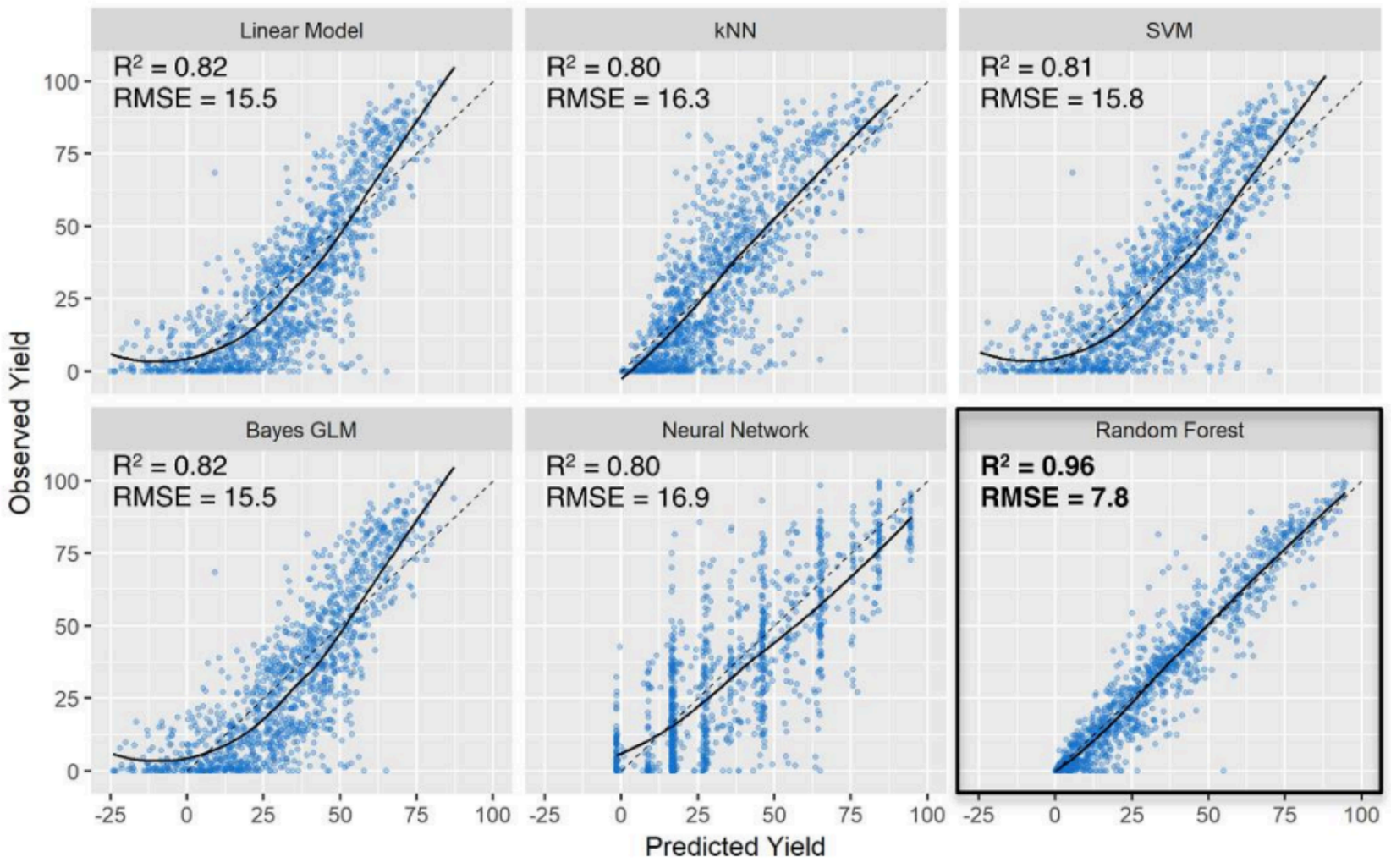
3 bases

- Ran 1536 well plates using a Mosquito robot
- Computationally pulled out molecular, atomic, and vibrational properties
- Examined linear regression and ML methods to try and predict the yield from the reaction parameters

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* ***2018****, 360, 186*.

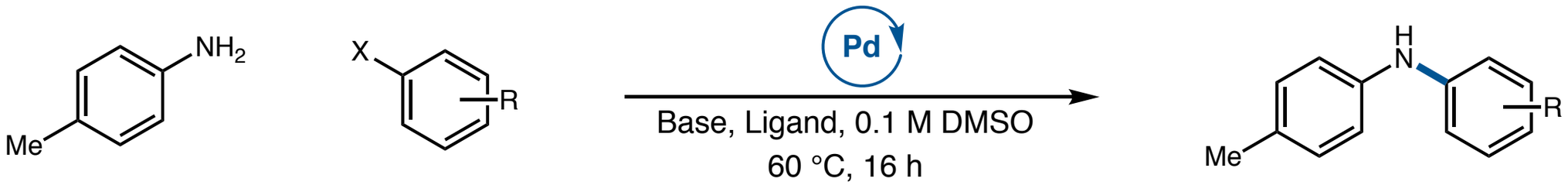
Predicting reaction performance

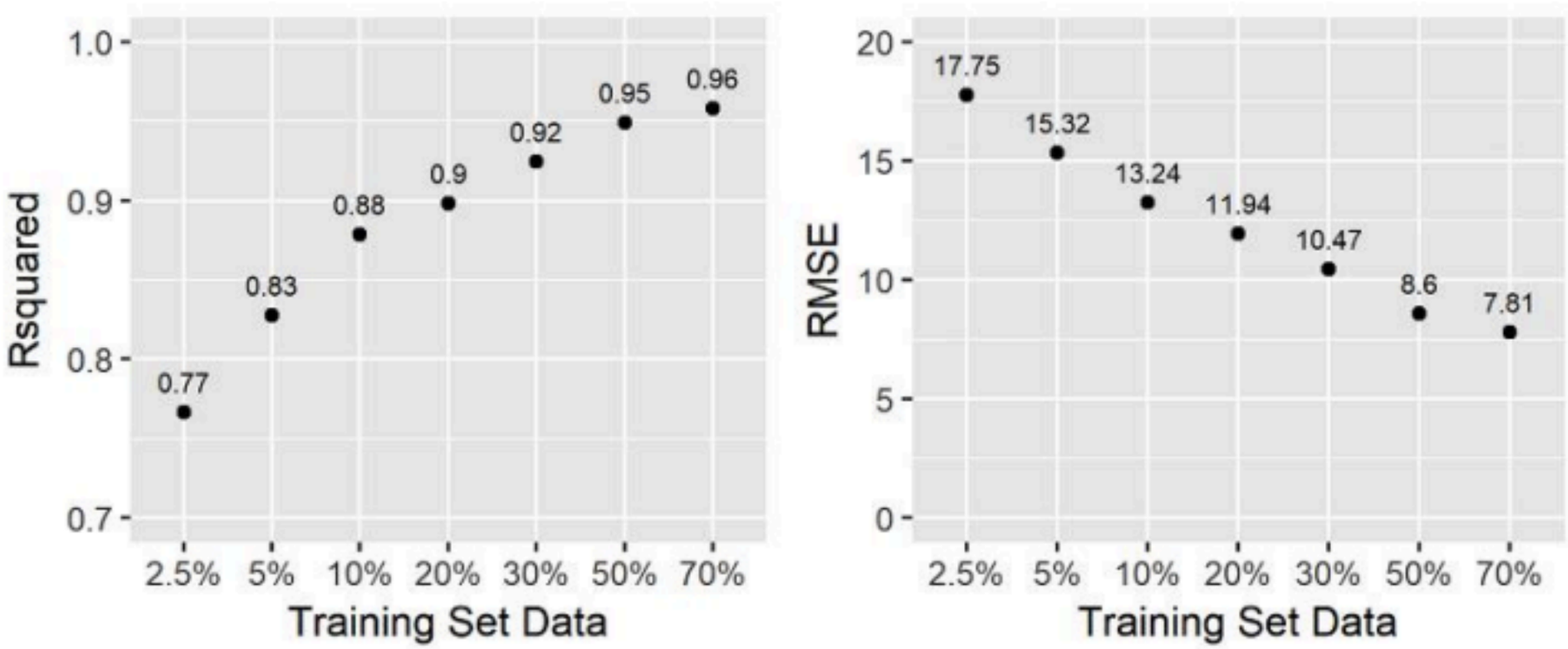




Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* ***2018****, 360, 186*.

Predicting reaction performance

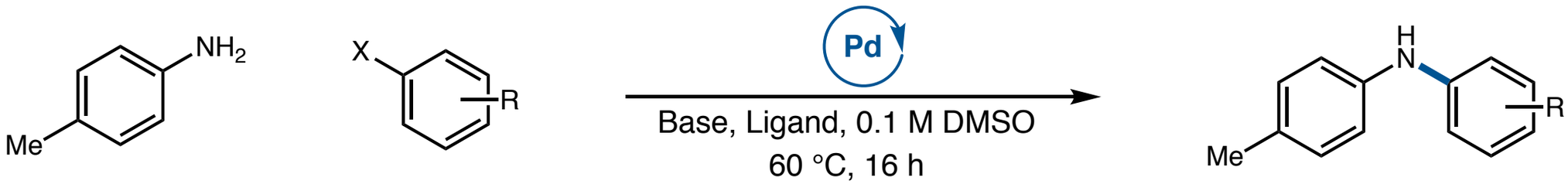




Examining training set data size

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* ***2018****, 360, 186*.

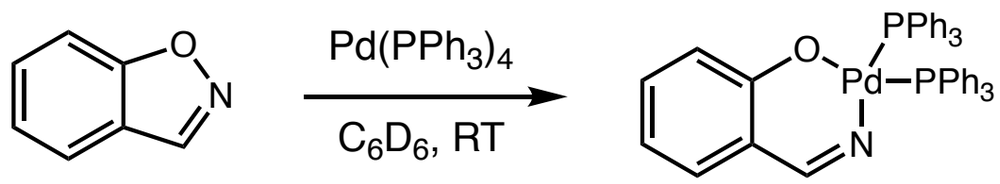
Predicting reaction performance



Most important reaction parameters

- Additive *C3 NMR shift
- Additive E LUMO
- Aryl halide *C3 NMR shift
- Additive *O1 electrostatic charge
- Additive *C5 electrostatic charge

Problems with oxidative addition into the isoxazole?



Oxidative addition product believed to be observed by ^{31}P, ^{1}H, and ^{13}C NMR

The observation of oxidative addition appeared to track well with the important reaction parameters

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* ***2018****, 360, 186*.

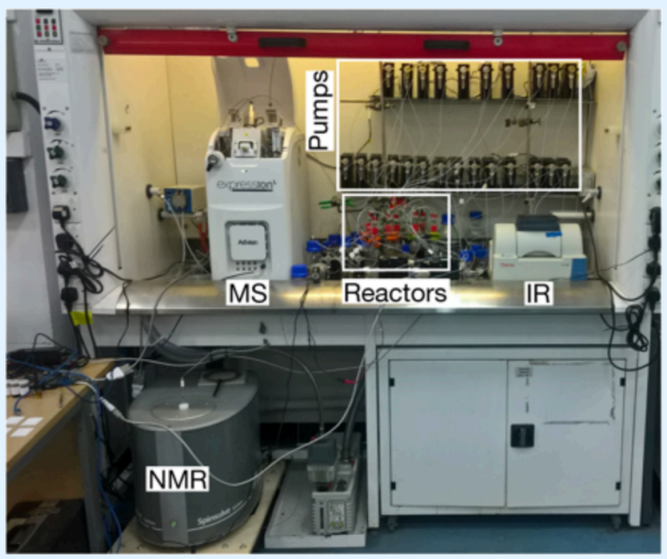
Reaction discovery using an automated robot

Can machine learning explore chemical space and predict areas of reactivity?

Examined different combinations of two- and three- component reactions from a pool of 18 starting materials using an automated robot

Automated Robot

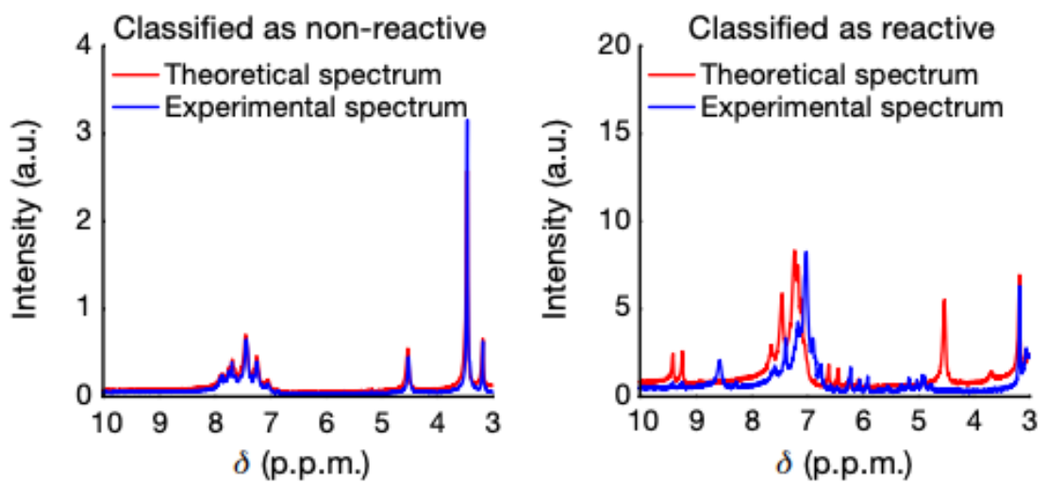
- Executes six experiments in parallel
- Analyzes reactions by ^{1}H NMR, MS, and IR



The reactions are classified as reactive or unreactive by a supported vector machine (SVM) with a linear kernel

This algorithm compared the IR and NMR spectra of the product to that of the starting material and registered differences as reactivity hits

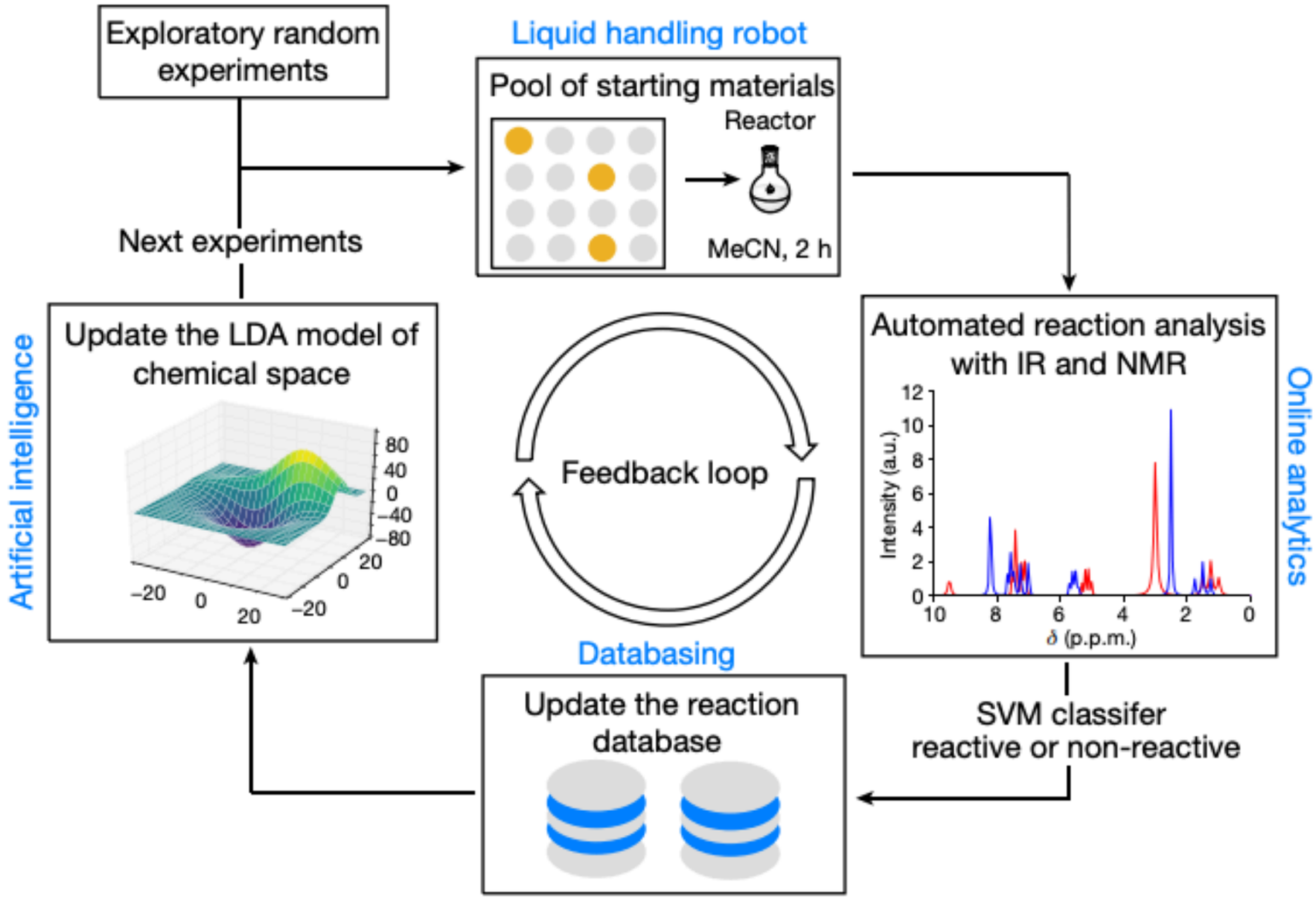
The SVM was trained on a set of 72 reactive and non reactive mixtures and could classify mixtures with an accuracy of 86%



SVM results were used to train a linear discriminant analysis (LDA) model that could construct a model of chemical space

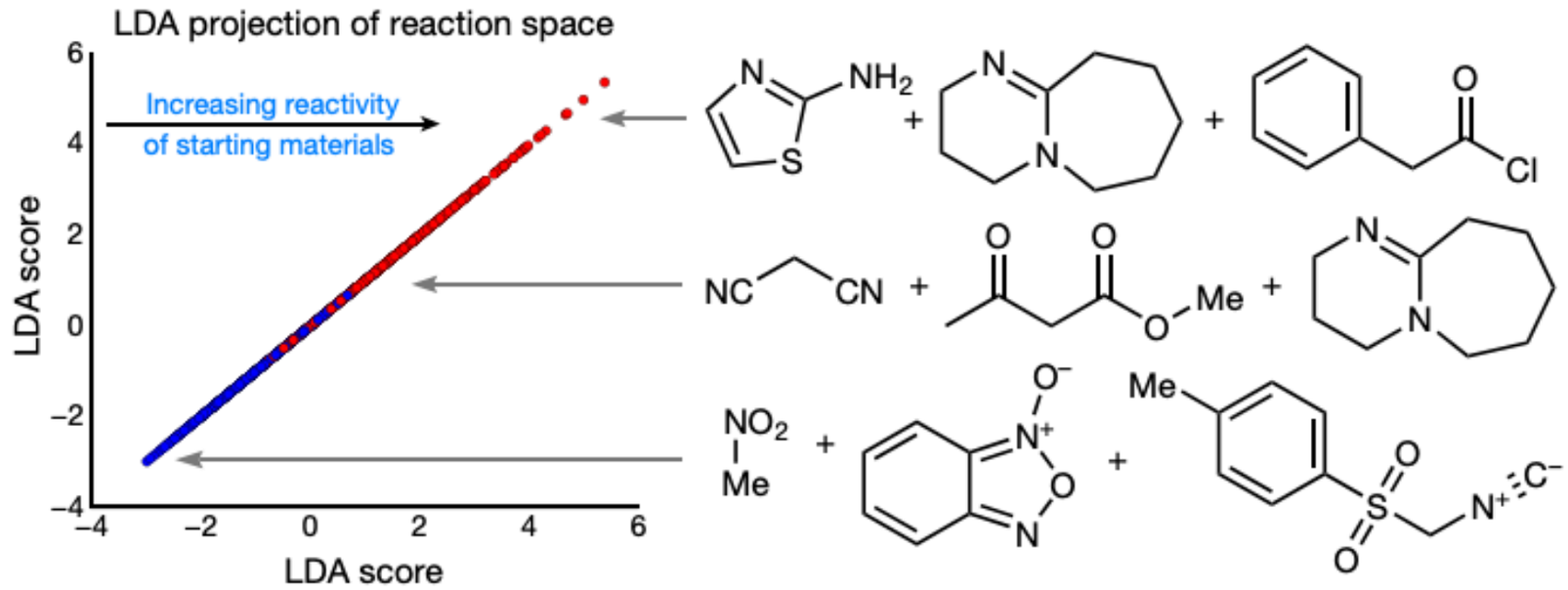
Granda, J. M.; Donina, L.; Dragone V.; Long, D.-L.; Cronin, L. *Nature* ***2018****, 559,* 377.

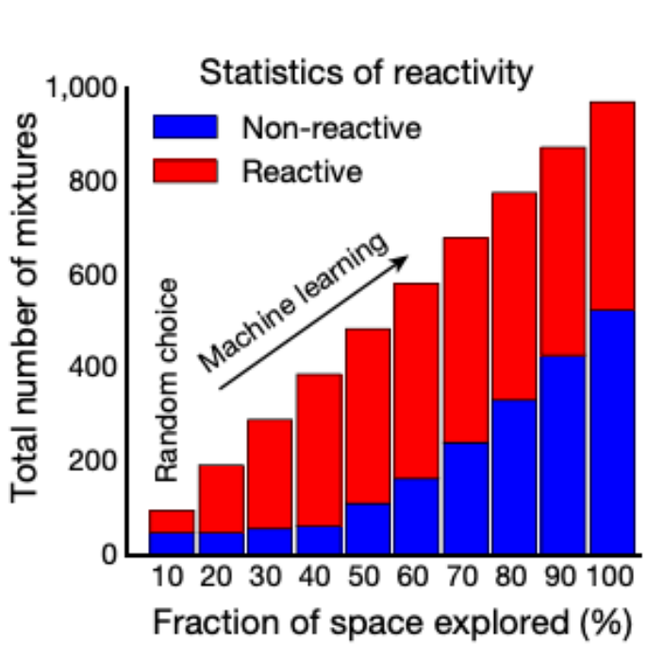
Reaction discovery using an automated robot

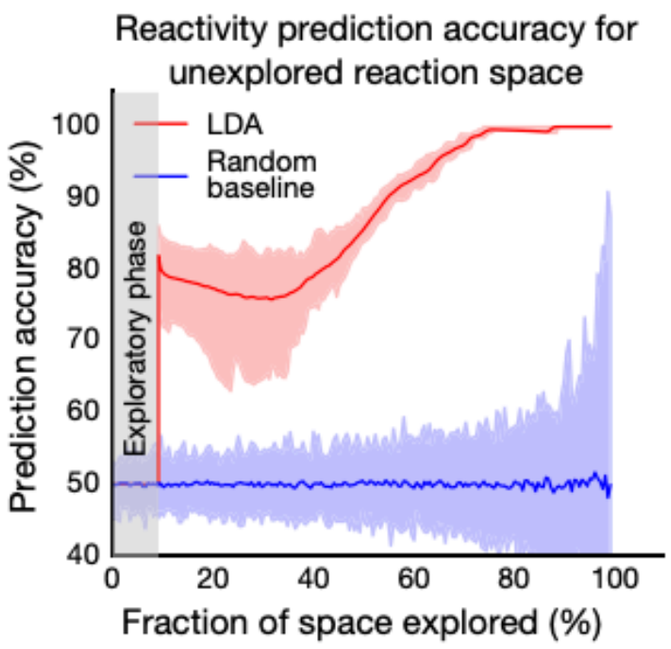


Granda, J. M.; Donina, L.; Dragone V.; Long, D.-L.; Cronin, L. *Nature* ***2018****, 559,* 377.

Reaction discovery using an automated robot





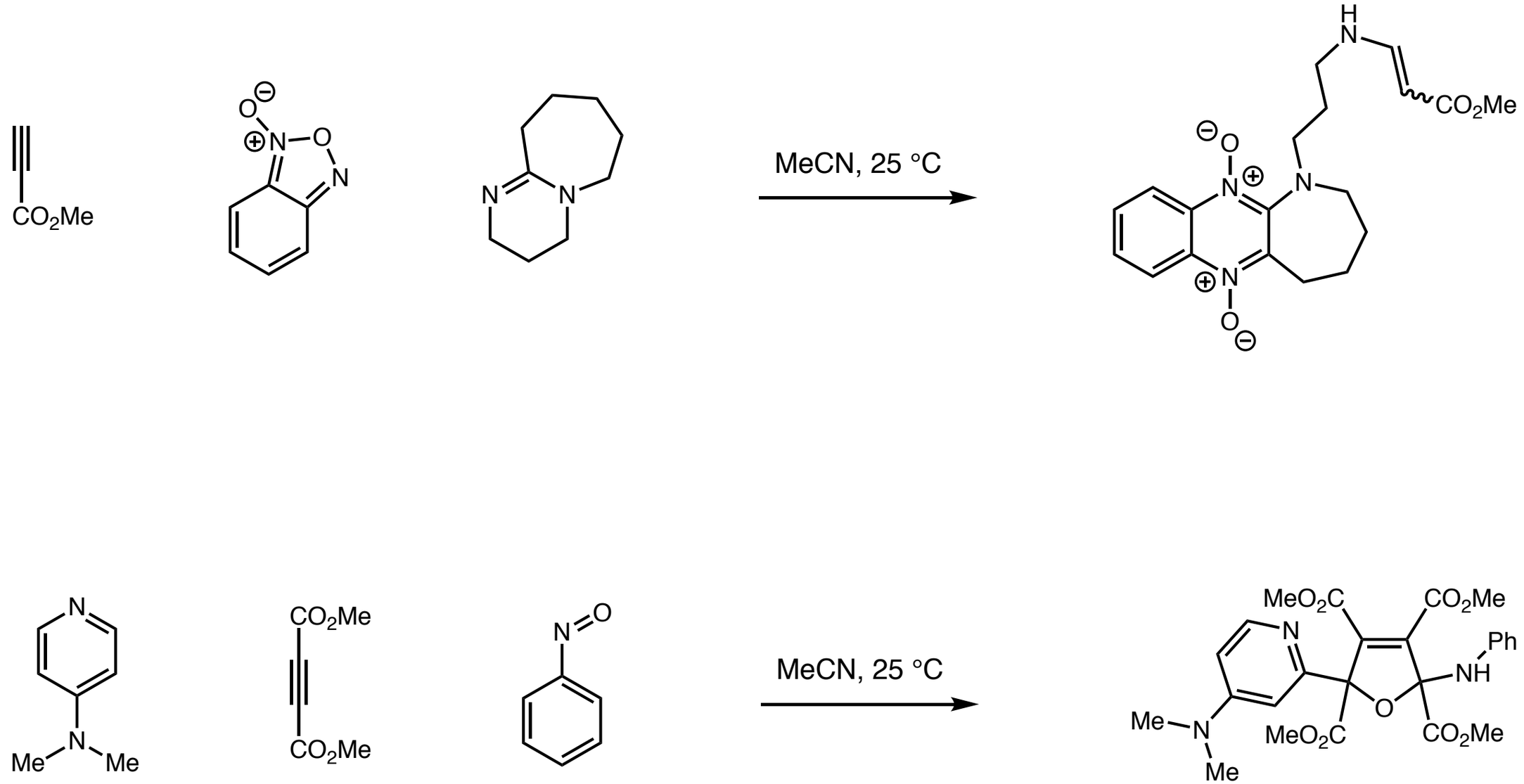


By surveying 10% of the reaction space, the model could accurately predict the reactivity of 80% of the chemical space

Granda, J. M.; Donina, L.; Dragone V.; Long, D.-L.; Cronin, L. *Nature* ***2018****, 559,* 377.

Reaction discovery using an automated robot

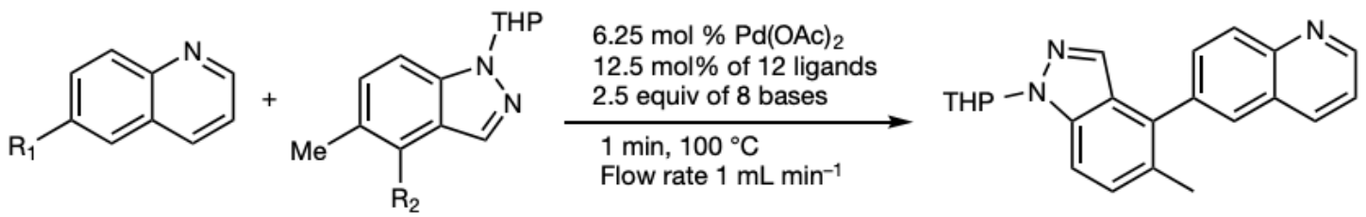
manual examination of the reactive combinations identified by the machine lead to the discovery of four novel transformations



Granda, J. M.; Donina, L.; Dragone V.; Long, D.-L.; Cronin, L. *Nature* ***2018****, 559,* 377.

Reaction discovery using an automated robot

This system could also be utilized for the prediction of yields



R_1 = Cl
R_1 = Br
R_1 = OTf
R_1 = I
R_1 = $B(OH)_2$
R_1 = Bpin
R_1 = BF_3K

R_2 = $B(OH)_2$
R_2 = Bpin
R_2 = BF_3K
R_2 = Br

Ligands

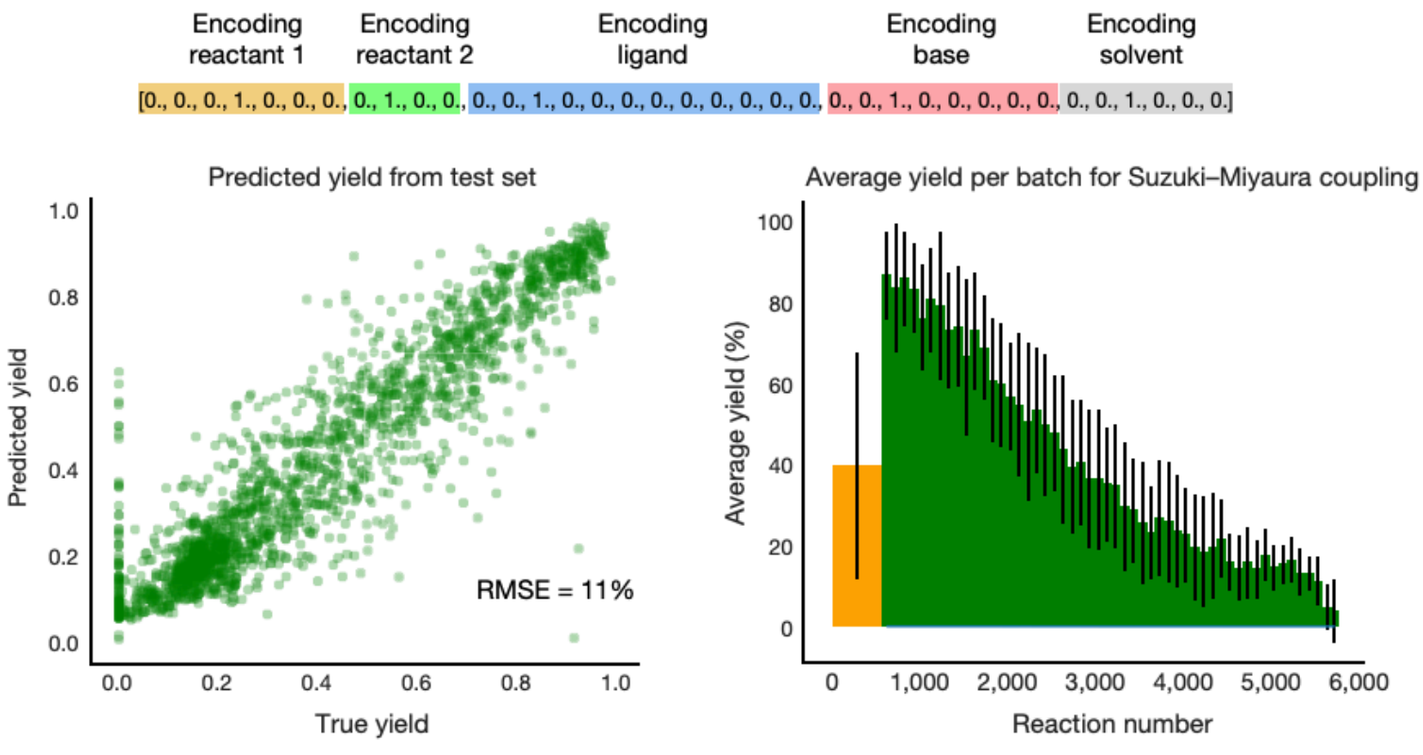
$P(tBu)_3$
$P(Ph)_3$
AmPhos
$P(Cy)_3$
$P(o\text{-}Tol)_3$
CataCXium
SPhos
dtbpf
XPhos
dppf
Xantphos
None

Bases

None
Et_3N
LiOtBu
CsF
K_3PO_4
KOH
$NaHCO_3$
NaOH

Solvents

$MeOH/H_2O$ 9:1
THF/H_2O 9:1
$MeCN/H_2O$ 9:1
DMF/H_2O 9:1

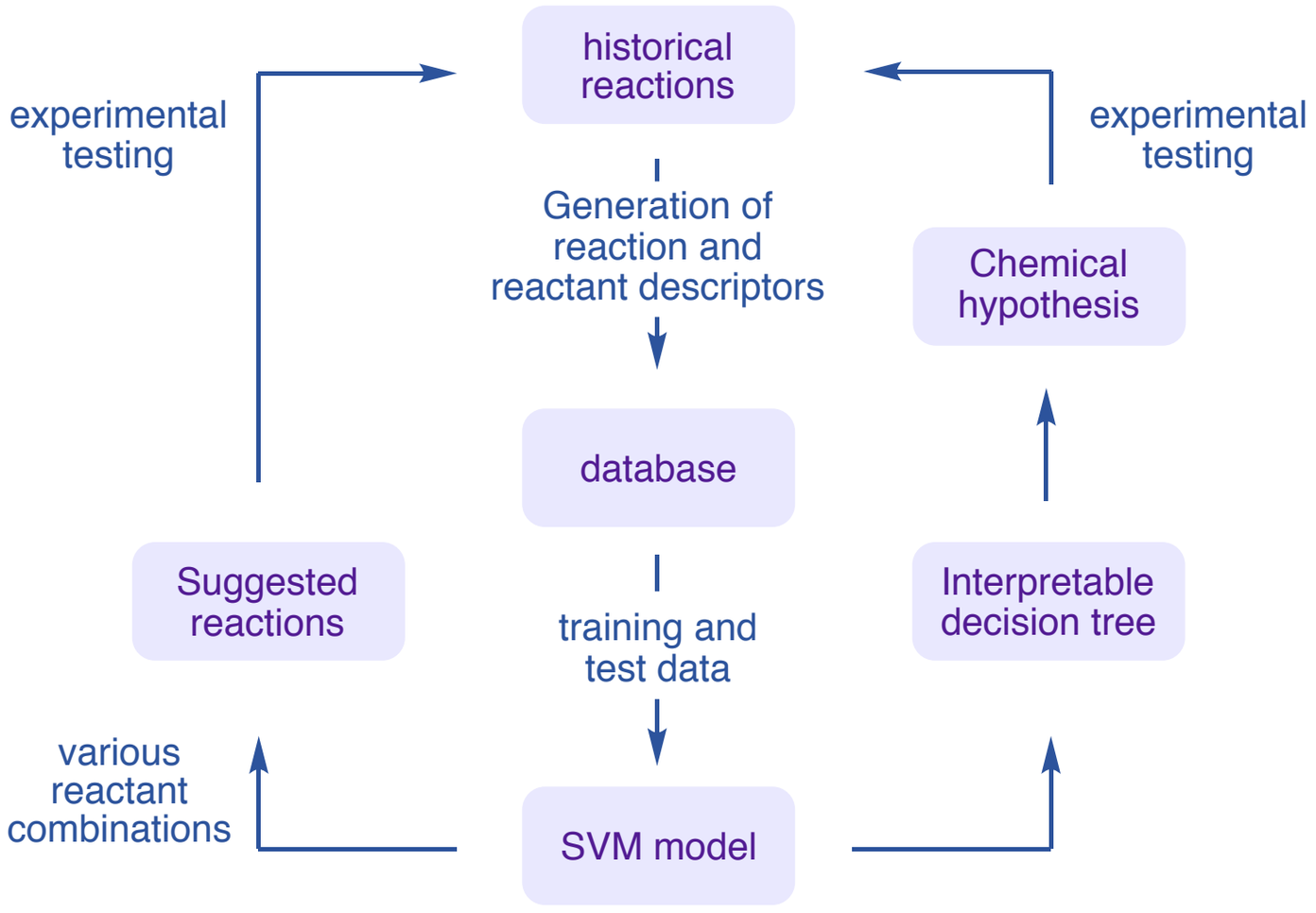


Granda, J. M.; Donina, L.; Dragone V.; Long, D.-L.; Cronin, L. *Nature* ***2018****, 559,* 377.

Applying machine learning to vanadium—selenite crystal synthesis

The synthesis of MOFs is not fully understood and development of new compounds relies primarily on trial and error

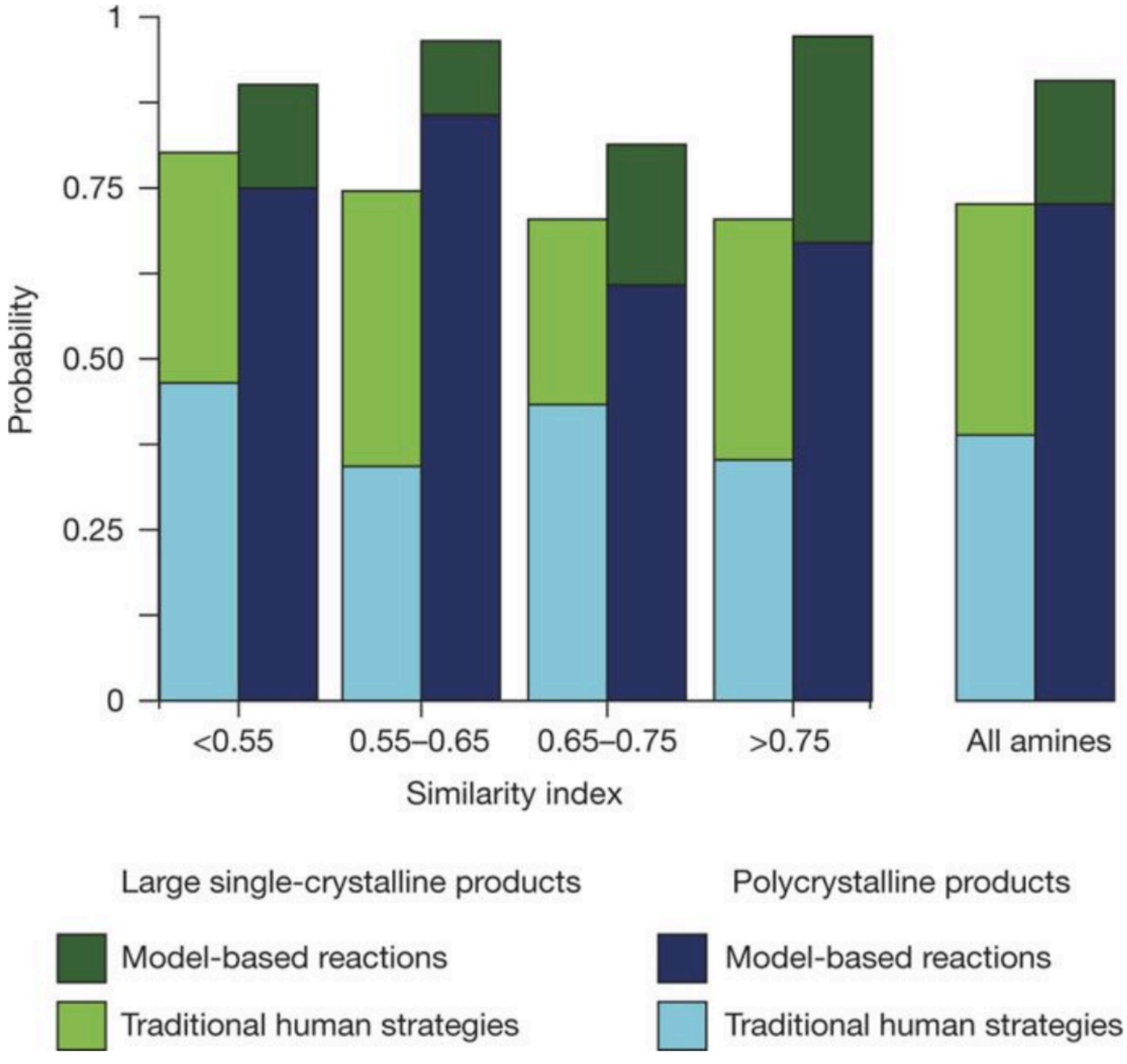
Acquired syntheses that failed from lab notebooks and utilized them to generate a chemical database



Raccuglia, P.; Elbert, K. D.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. *Nature* **2016**, *533*, 73.

Applying machine learning to vanadium—selenite crystal synthesis

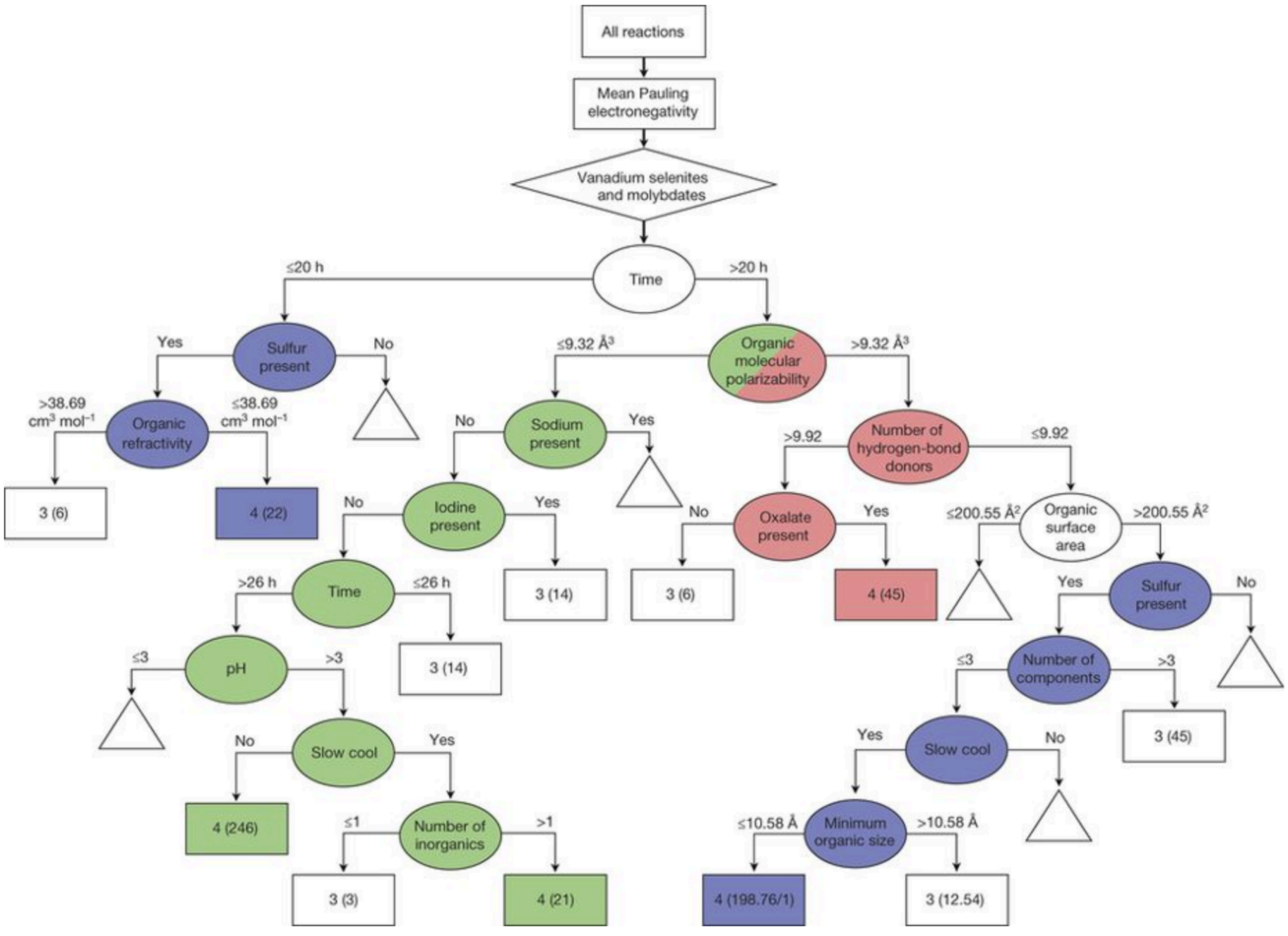
The model outperformed traditional human strategies by a statistically significant margin



Raccuglia, P.; Elbert, K. D.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. *Nature* **2016**, *533*, 73.

Applying machine learning to vanadium—selenite crystal synthesis

The SVM model was converted to a decision tree to make it chemically interpretable



Raccuglia, P.; Elbert, K. D.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. *Nature* **2016**, *533*, 73.

Applying machine learning to vanadium—selenite crystal synthesis

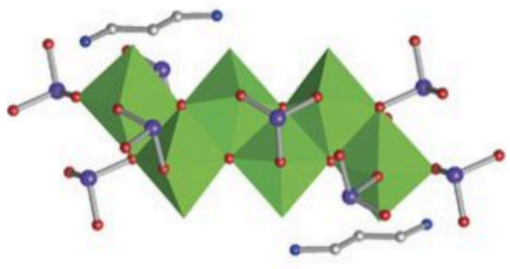
This decision tree could be utilized to make chemical hypotheses

Small, low-polarizability amines

↓

No Na^+, longer reaction times

1,3-Diaminopropane



$[C_3H_{12}N_2][V_3O_5(SeO_3)_3]\cdot H_2O$

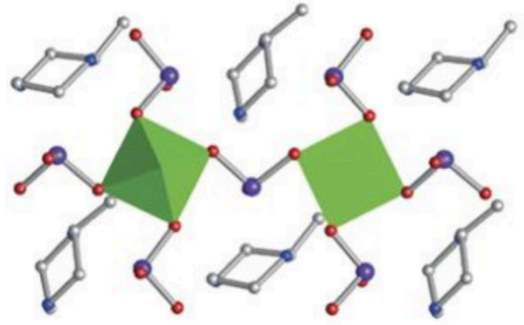
- Avoid preciptating inorganic building units

Spherical, low-projection-size amines

↓

V^{4+} reagents required

2-Methylpiperazine



$[C_5H_{14}N_2][VO(SeO_3)_2]$

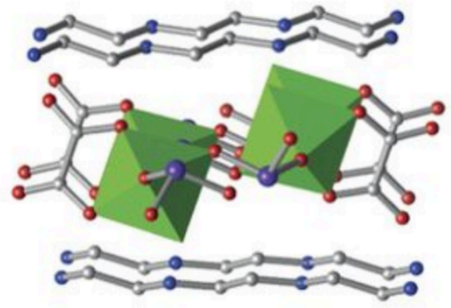
- Unable to directly generate V^{4+} *in situ* from V^{5+}

Long, linear tri- and tetramines

↓

Oxalate component required

Triethylenetetramine



$[C_6H_{22}N_4][VO(C_2O_4)(SeO_3)]_2\cdot 2H_2O$

- Alter the charge density of inorganic secondary building units

Raccuglia, P.; Elbert, K. D.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. *Nature* **2016**, *533*, 73.

Machine learning

Retrosynthesis

Reaction optimization

- Identifying ideal condtions
- Predicting yields

Medicinal Chemistry

- Structure generation
- Binding prediction
- Estimating ADMET properties

Chemistry

Reaction discovery

Materials chemistry

- Identifying ideal alloys
- Synthesis/ Crystalization

Questions?