The Dark Reactions Project: A Case Study in Using Machine Learning to Accelerate Exploratory Materials Synthesis

Joshua Schrier

Haverford College / Fordham University

A Related Effort:

The Mission Innovation *Materials Acceleration Platform* Workshop Report (Jan 2018)

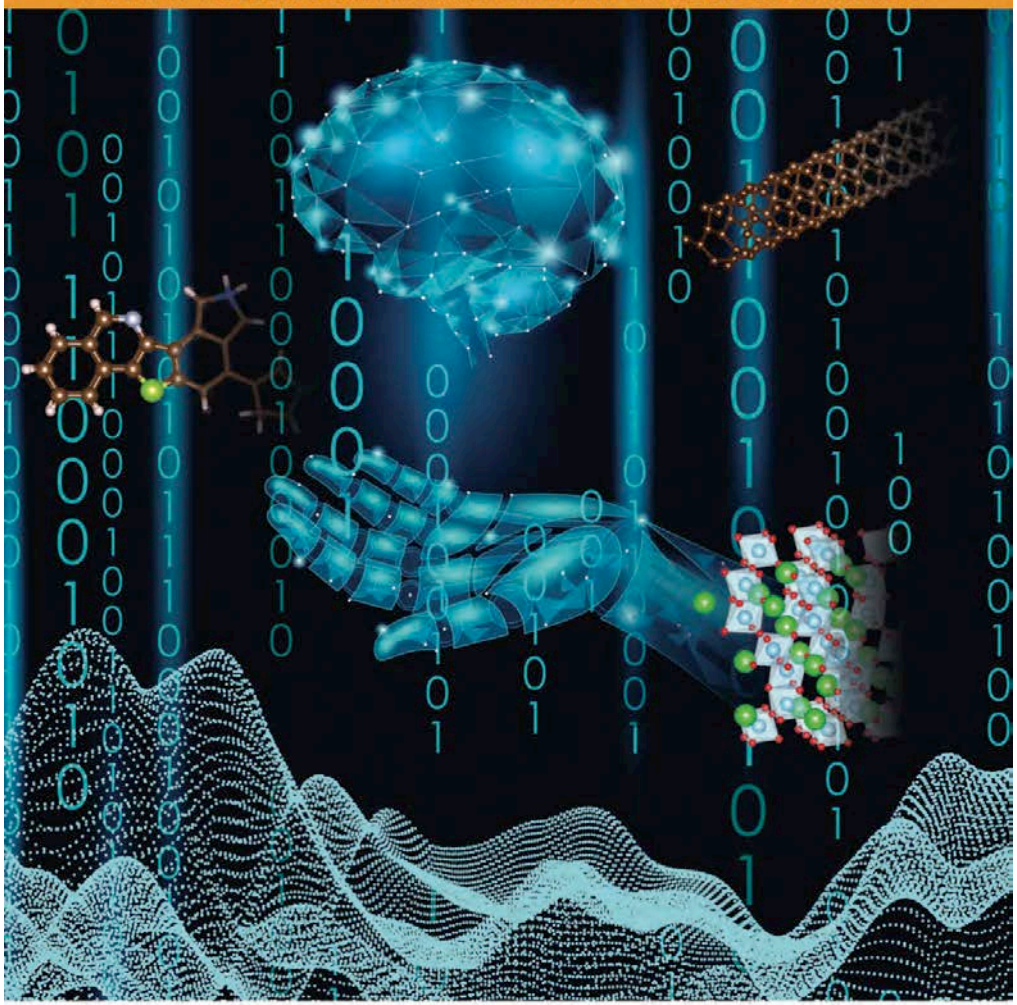
MATERIALS ACCELERATION PLATFORM

Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods with Artificial Intelligence

Mission Innovation

January 2018

REPORT OF THE CLEAN ENERGY MATERIALS INNOVATION CHALLENGE EXPERT WORKSHOP



SENER
SECRETARÍA DE ENERGÍA

U.S. DEPARTMENT OF ENERGY

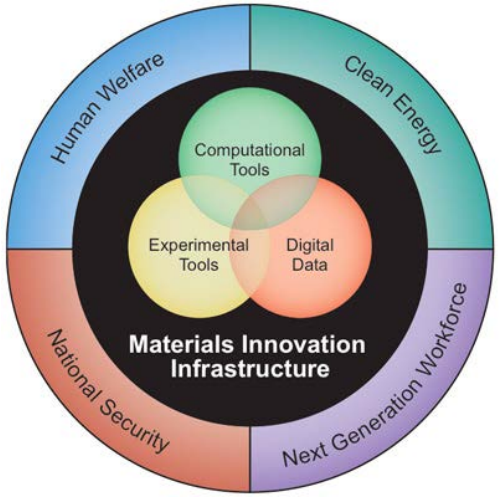
CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

The Workshop

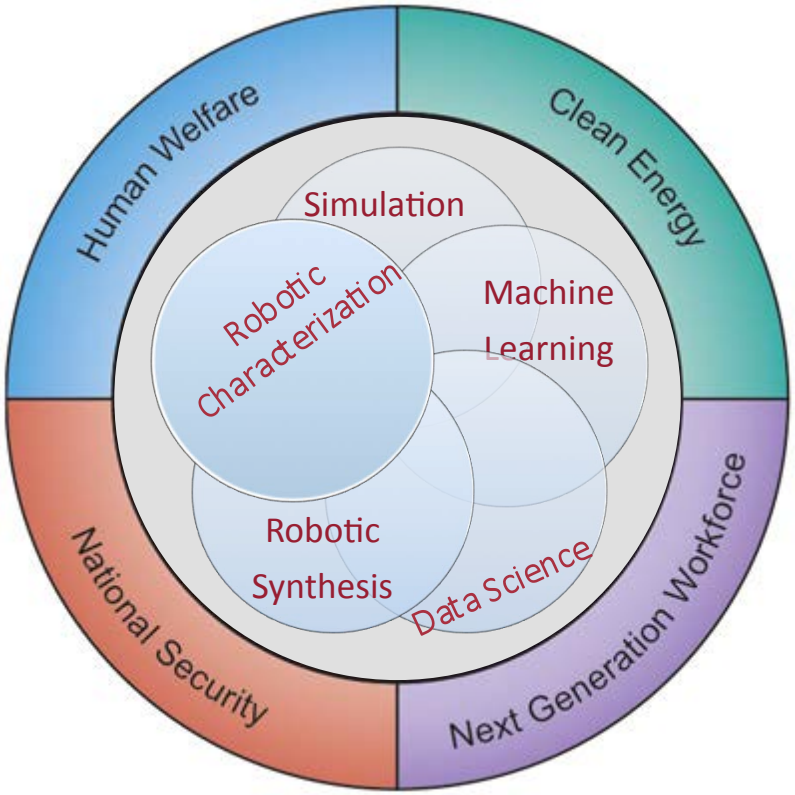
The workshop (September 2017 in Mexico City) drew 133 attendees:

- 55 professors and scientists from top universities and research institutions;
- 6 keynote speakers and panellists, including Nobel Laureate Dr. Mario Molina;
- 16 MI member governments represented: Australia, Canada, Denmark, Finland, France, Germany, European Union, India, Italy, Korea, Mexico, Netherlands, Norway, Saudi Arabia, United Kingdom, and United States;
- affiliates of Mexico- and U.S.-based universities, groups, labs, and companies;
- graduate students and postdoctoral researchers; and observers from different countries

Updating the vision of MGI 1.0

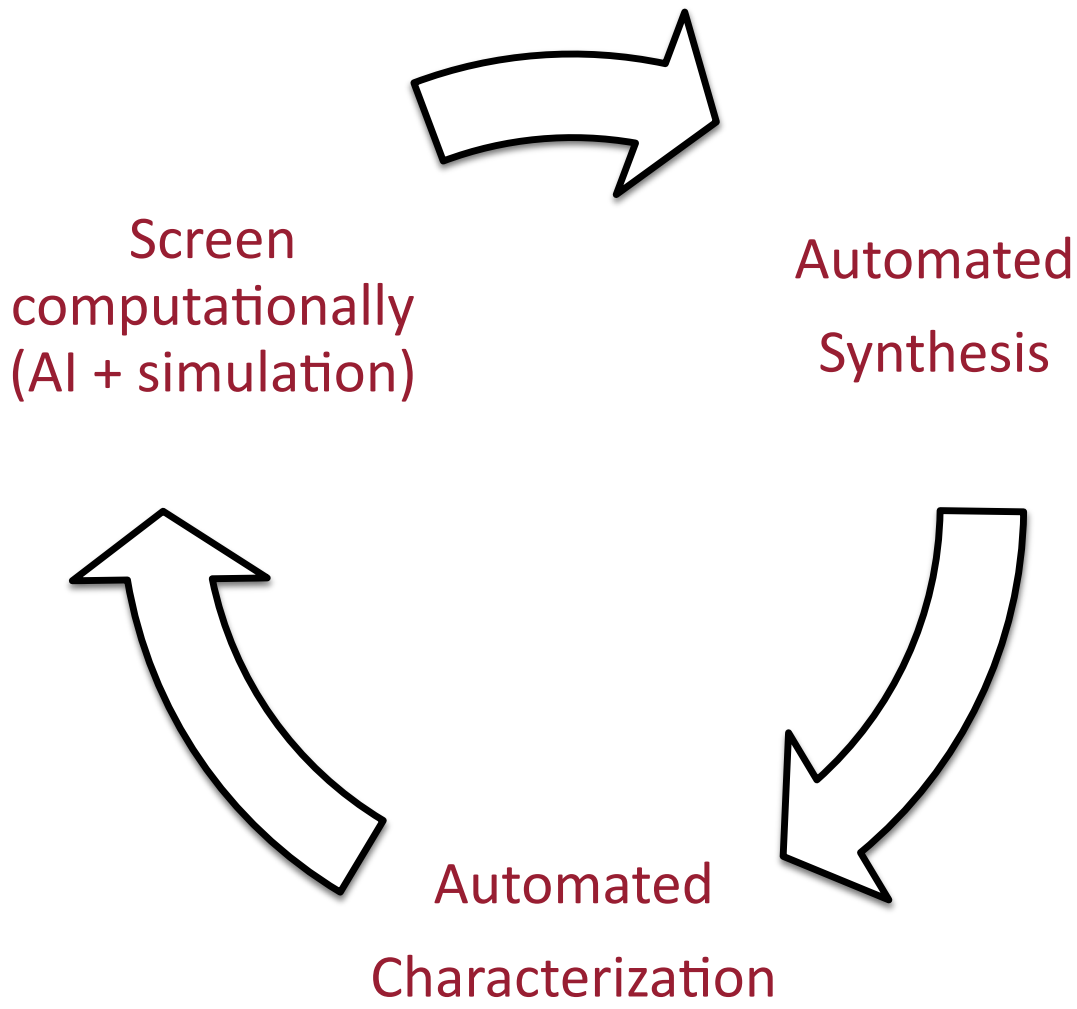


Materials Genome Initiative
2011

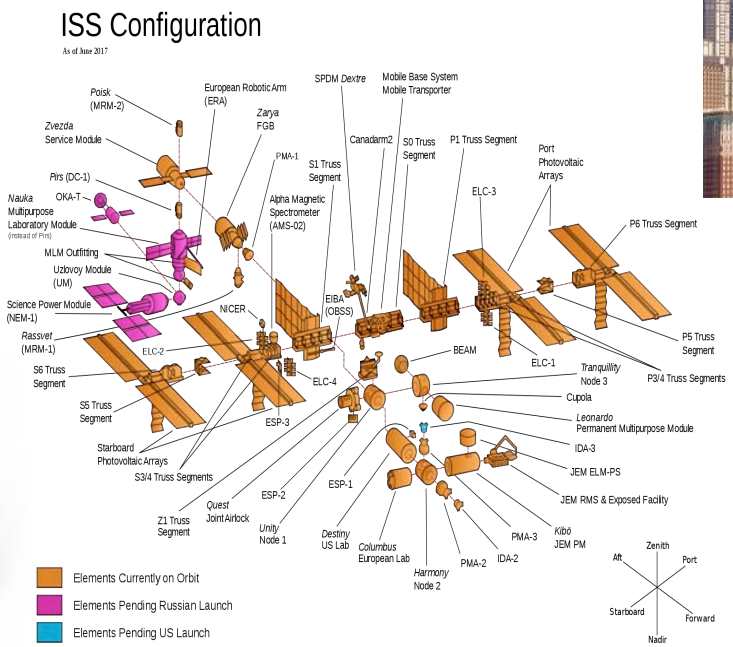


Mission Innovation Challenge 6
2017

Main Recommendation: Close the Loop!



Biggest surprise: Importance of modularity.



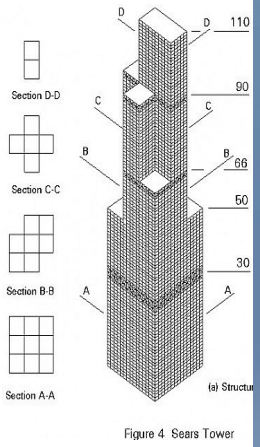


Figure 4 Sears Tower

Sears Tower

Burj Khalifa

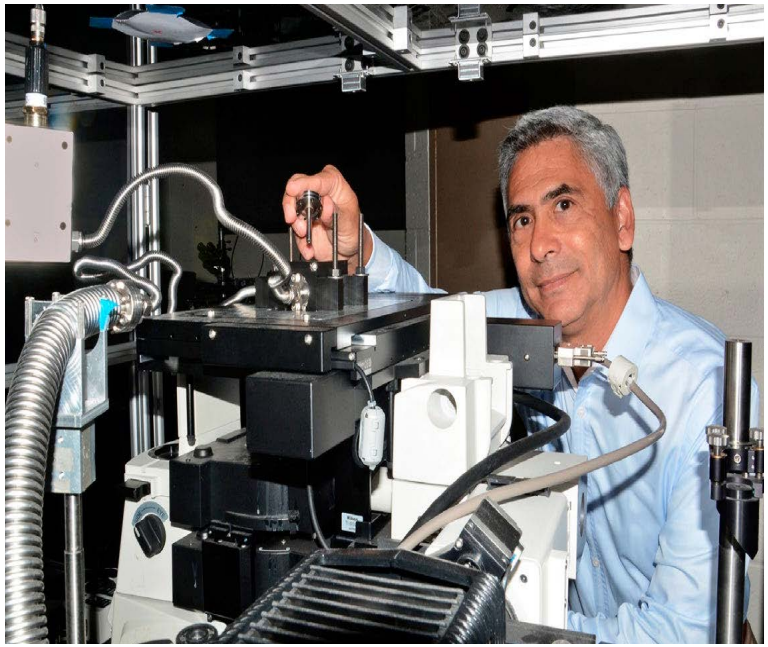
MAP elements

1. Closing the loop
2. AI for Materials
3. Modular Materials Robotics
4. Inverse Design
5. Bridging Length and Timescales
6. Data Infrastructure and Exchange

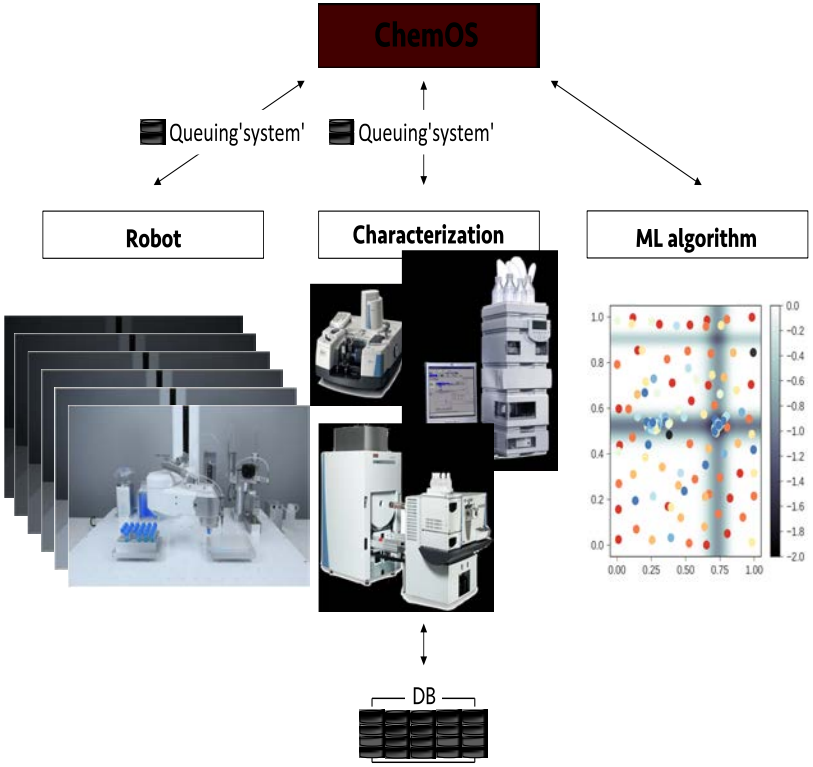
1. Closing the loop

Integrate powerful, yet usually separate elements of materials design in a closed loop

Autonomous closed-loop nanoparticle synthesis system
Benji Maruyama
Air Force Research Laboratories



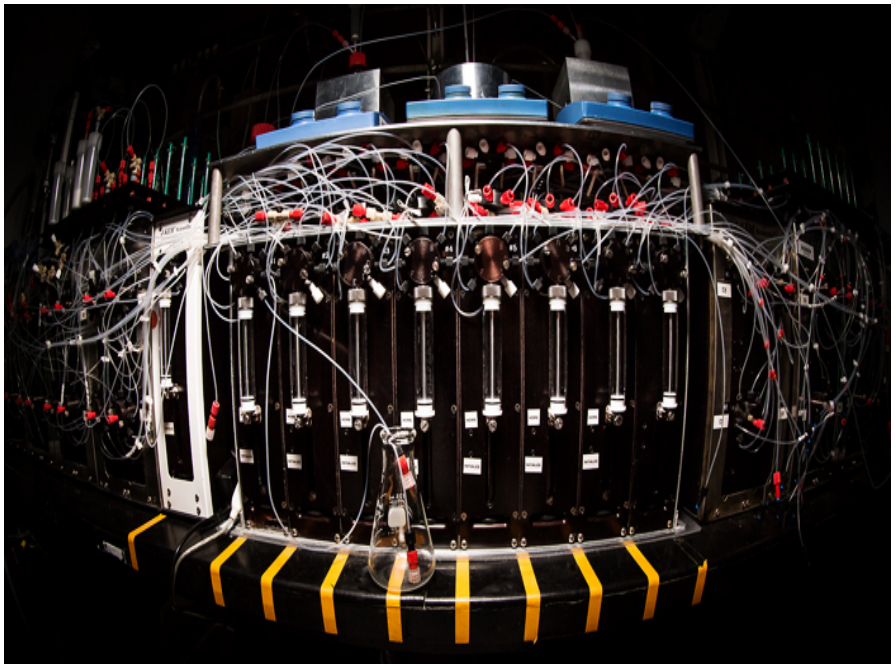
2. AI for Materials



The scale and details of theoretical, computational, synthetic, and characterization evidence in materials research require the establishment of this new branch of AI.

National and international research organizations must facilitate an integrated computer and materials science research effort to develop algorithms that mimic, and then supersede, the intellect and intuition of expert materials scientists.

3. Modular Materials Robotics



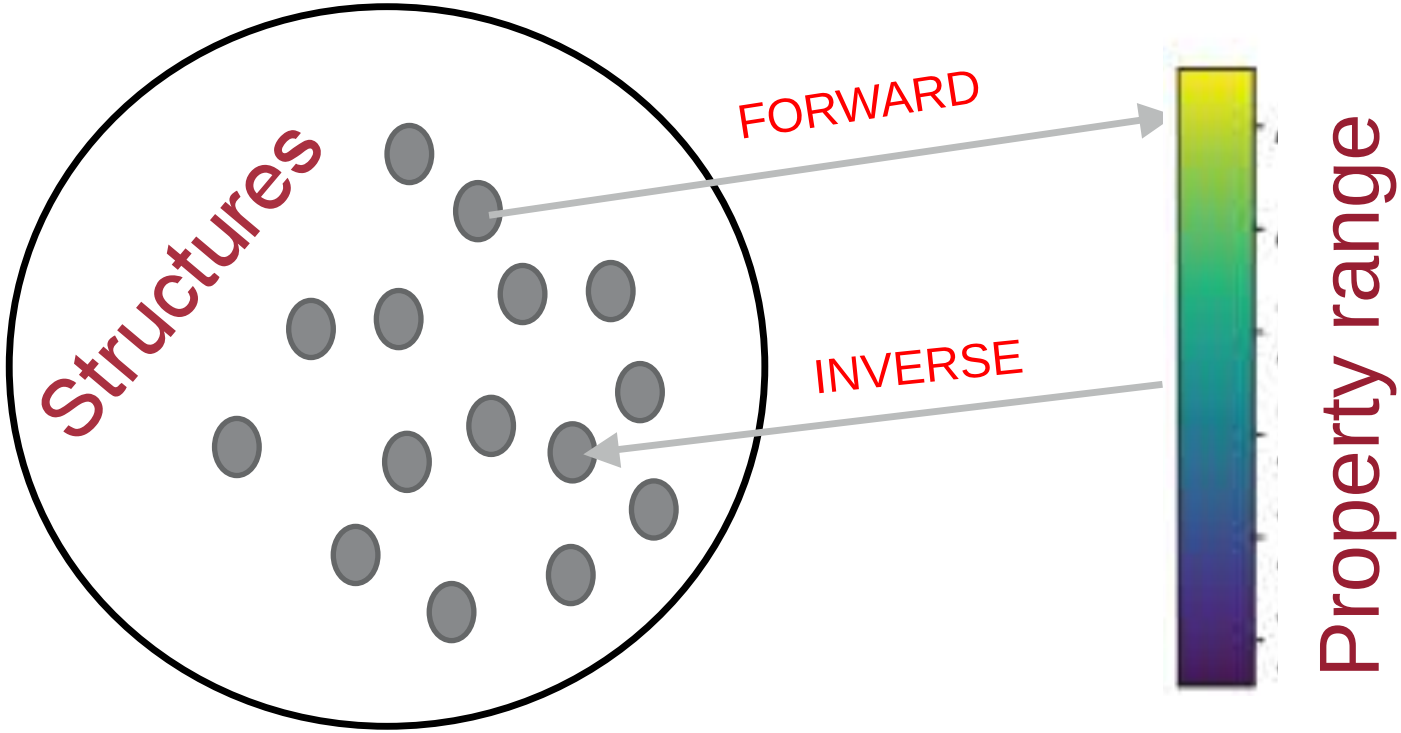
The Synthesis Machine
Marty Burke, University of Illiniois at Urbana Champaign

To accommodate evolving materials demands and the ever-expanding breadth of clean energy technologies, autonomous laboratories must remain nimble and motivate a modular approach to the development of materials science automation.

Treating techniques and materials as *modular building blocks* fosters human-machine communication and simplifies the path to materials exploration beyond the bounds of known materials.

4. Inverse Design

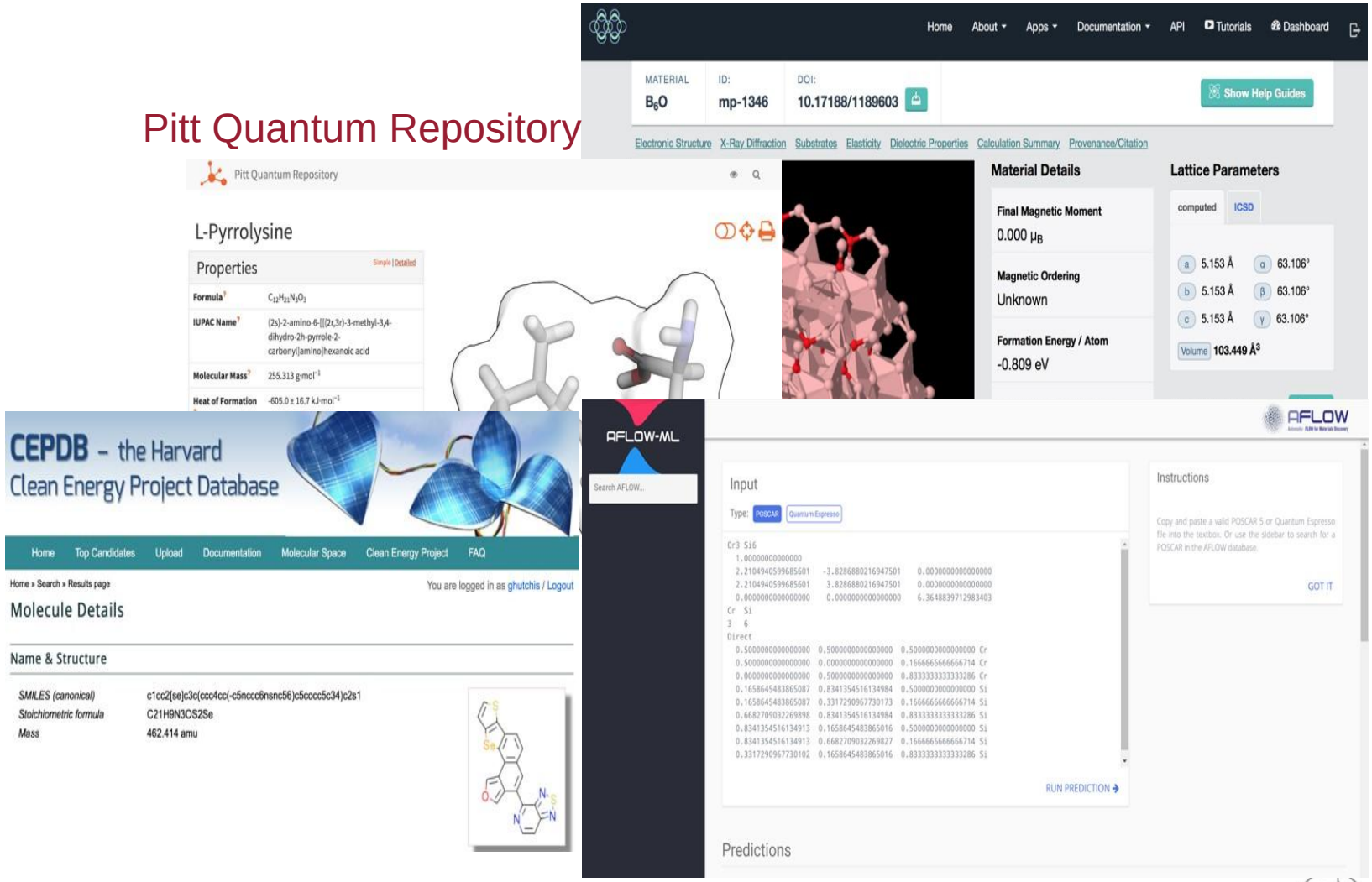
Inverse design enables automated generation of candidate materials designed to meet the performance, cost, and compatibility requirements of a given clean energy technology.



6. Data Infrastructure

The Materials Project

Pitt Quantum Repository



Harvard Clean Energy Project

AFLOW-lib and AFLOW-ml

HAVERFORD COLLEGE

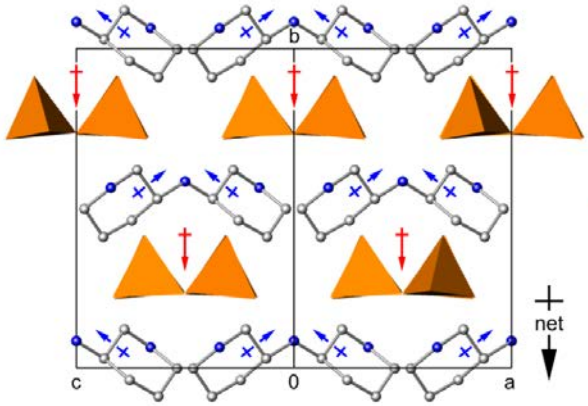
So where are we today?

A case study on
“The Dark Reactions Project”

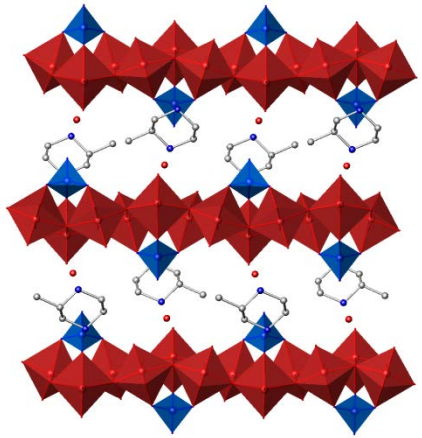
Initial Motivations

- Most synthesis reactions are "failures"
  - Archived in laboratory notebooks
  - Never reported in the literature
  - "Dark reactions"
- Can we learn from the "dark reactions"?
  - Define boundary of "successful" reactions
  - Use as a dataset for machine learning
  - Predict new reactions? Speed up discovery?
  - Develop new (testable) chemical hypotheses?

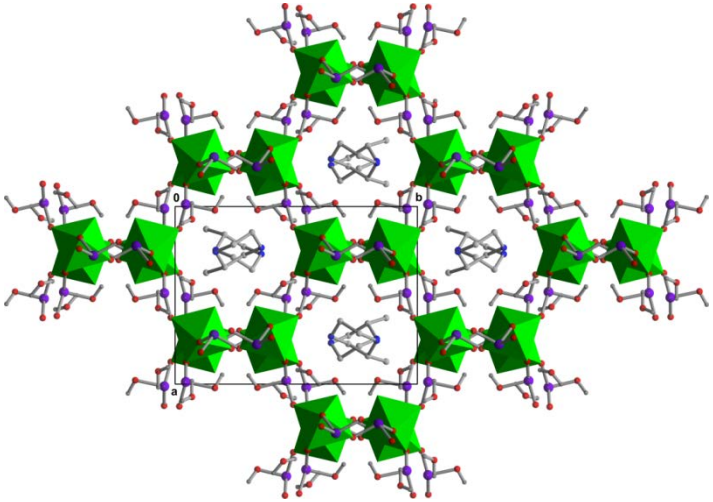
Amine-Templated Metal Oxides…



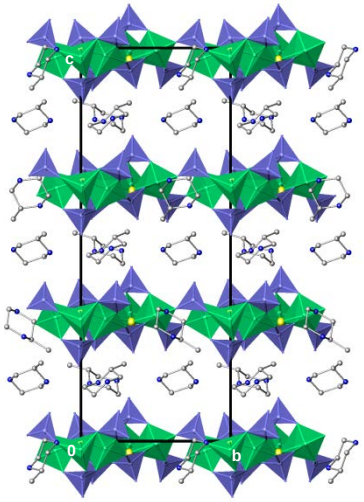
Norquist *et al. J. Solid State Chem.* **2012,** *195*, 86.



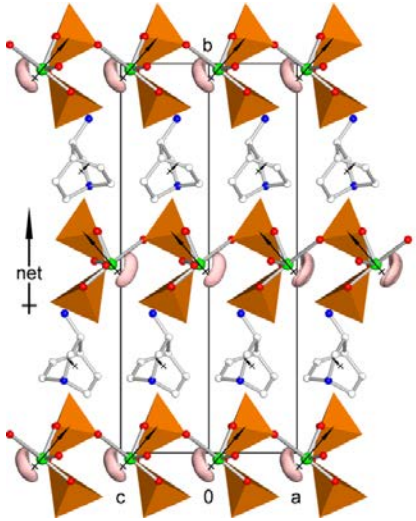
Norquist *et al. Inorg. Chem.* **2004,** *43*, 6528;
Norquist *et al. Cryst. Growth Des.* **2005,** *5*, 1913.



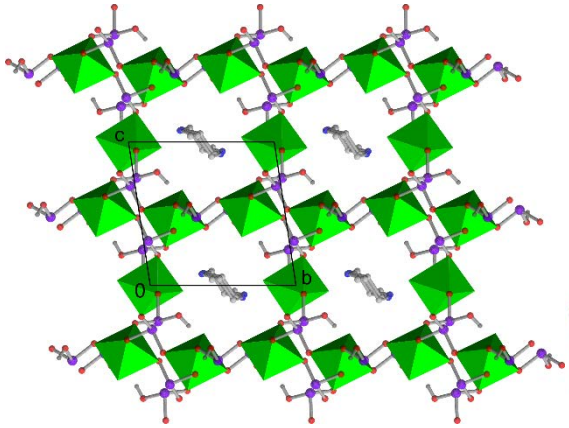
Norquist *et al. Inorg. Chem.* **2012,** *51*, 11040.



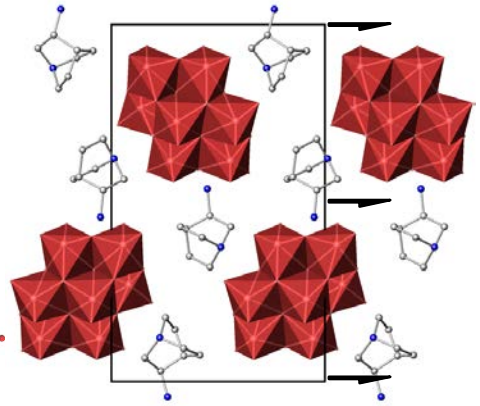
Norquist *et al. Inorg. Chem.* **2009,** *48*, 11277.



Norquist *et al. J. Solid State Chem.* **2011,** *184*, 1445.

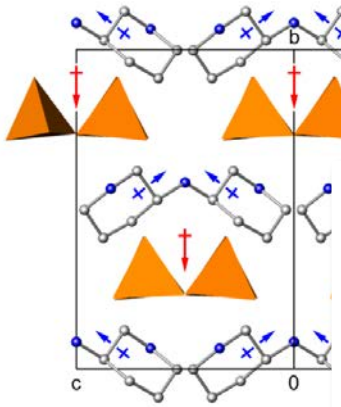


Norquist *et al. Inorg. Chem.* **2014,** *53*, 12027.

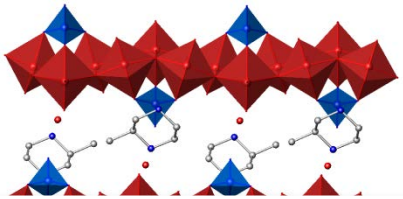


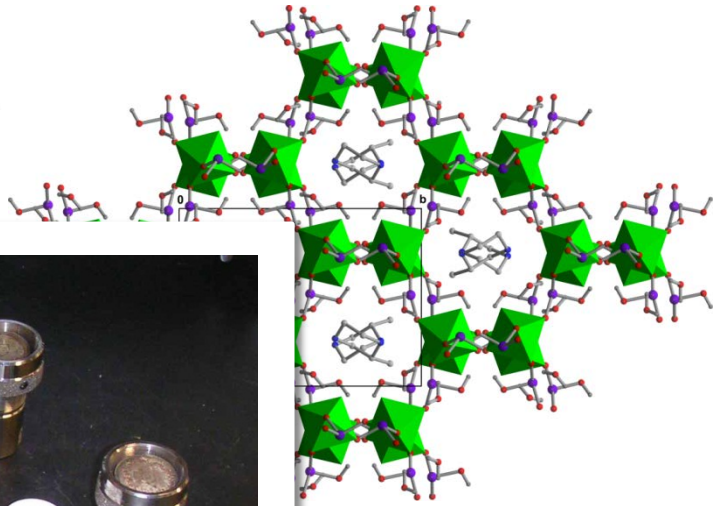
Norquist *et al. Inorg. Chem.* **2006,** *45*, 5529.

...produced by hydrothermal synthesis.

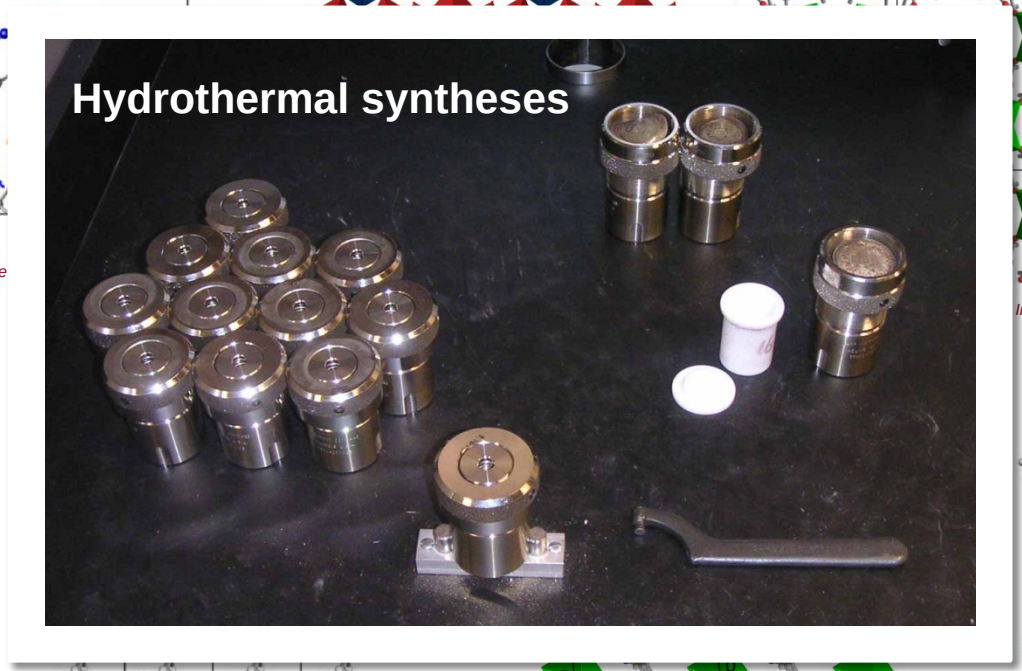


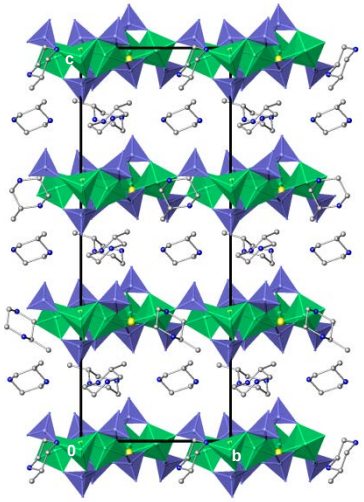
Norquist *et al. J. Solid State*



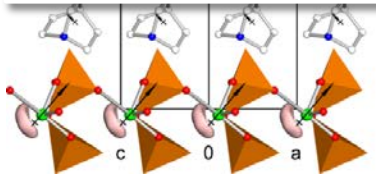


Inorg. Chem. **2012,** *51*, 11040.

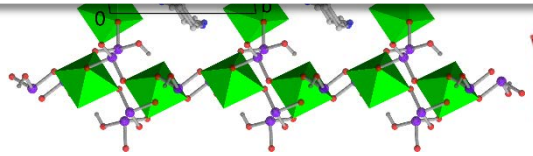




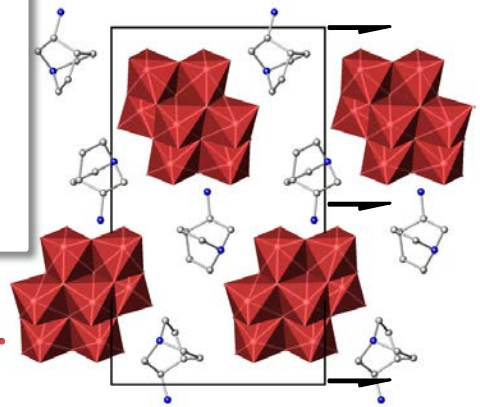
Norquist *et al. Inorg. Chem.* **2009,** *48*, 11277.



Norquist *et al. J. Solid State Chem.* **2011,** *184*, 1445.



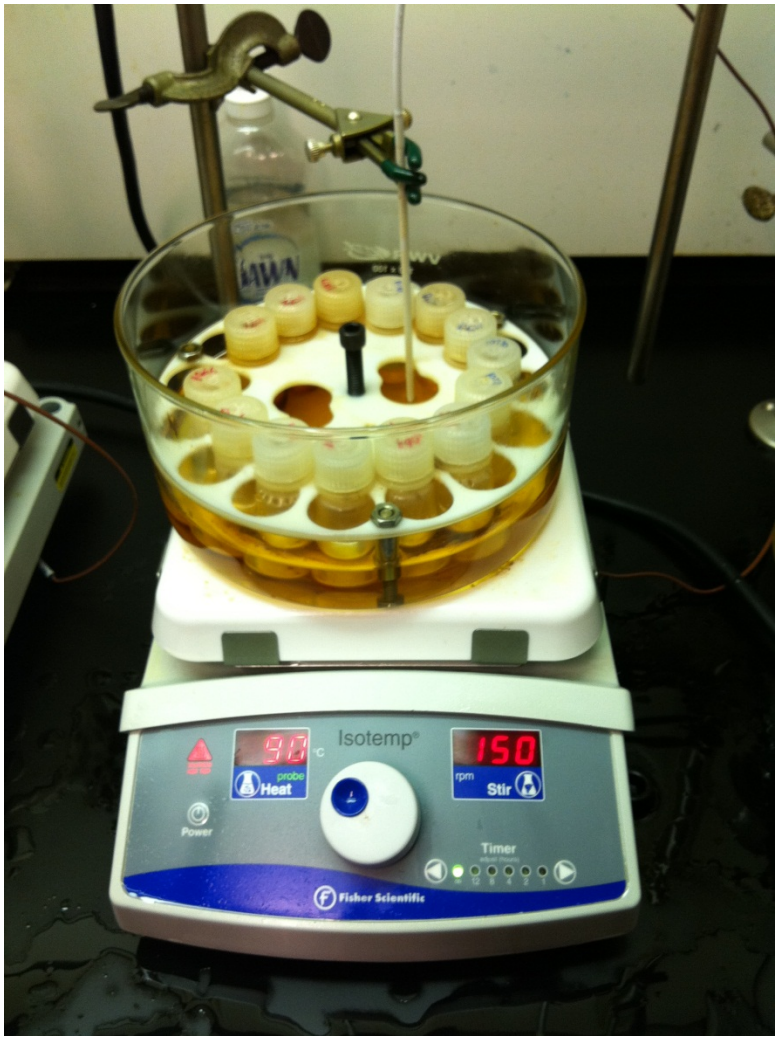
Norquist *et al. Inorg. Chem.* **2014,** *53*, 12027.



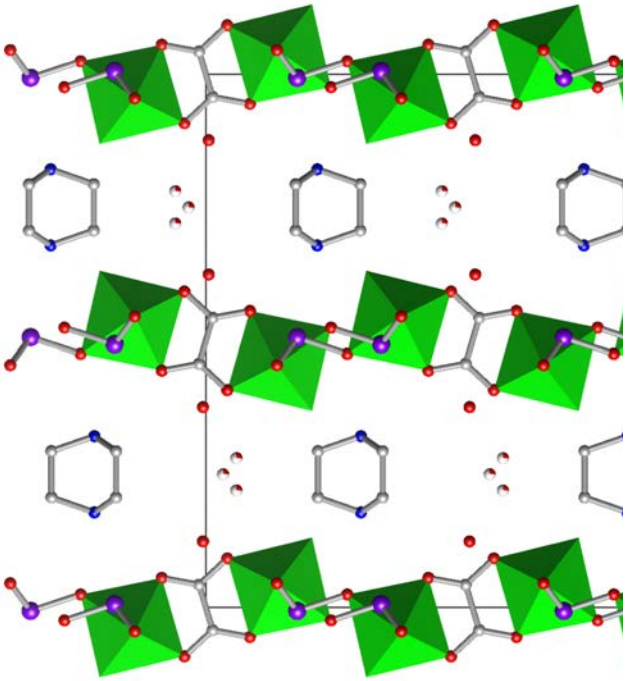
Norquist *et al. Inorg. Chem.* **2006,** *45*, 5529.

...described by a simple, *modular* recipe.

- Two metal sources
- One amine
- (Oxalic acid)
- Water
- pH
- Temperature
- Time

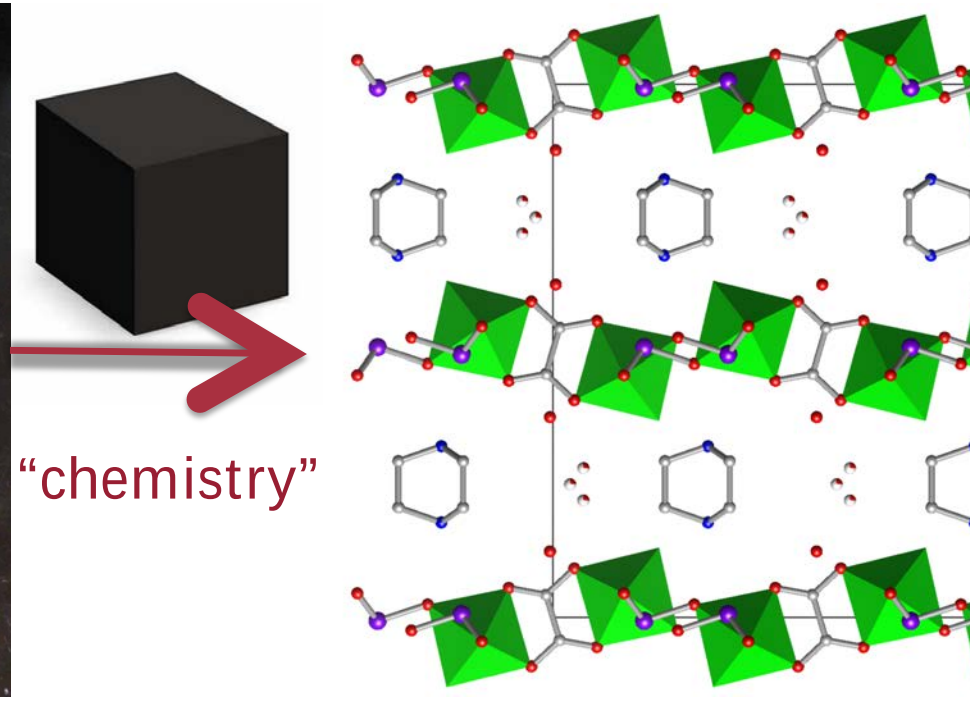


The Problem

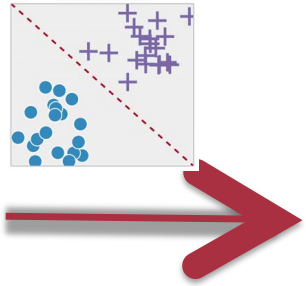


- Exploratory synthesis—vast chemical space
  - Can <u>not</u> predict reaction outcome!
    - Unanticipated structures are often observed
    - No predictive model for reaction success

Is this simply a classifier problem?



<u>Inputs:</u>
Concentration
pH
Time
Temperature
Etc.



"machine learning"

<u>Outputs:</u>
Big crystal
Small crystal
Something bad ("tar")
No reaction

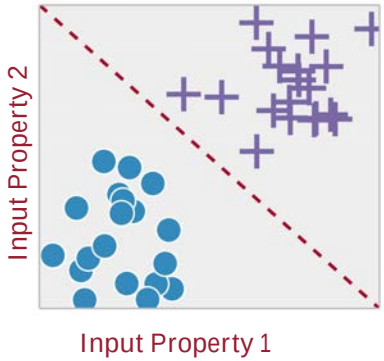
Including data on unpublished "failed" experiments is essential for reliable ML models.

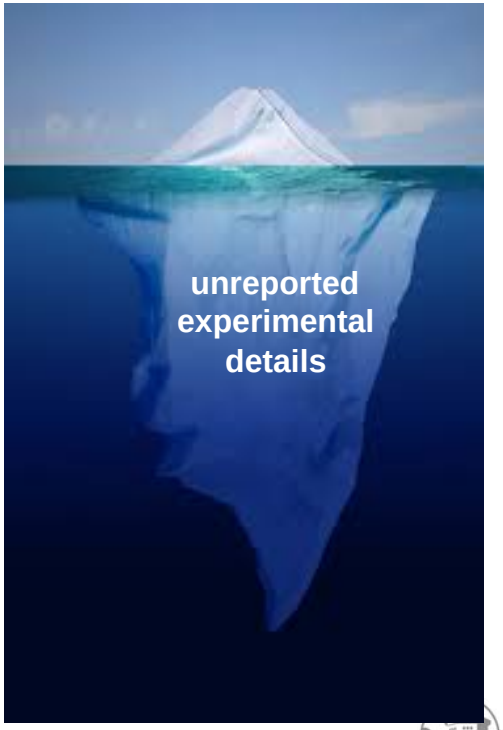
- To distinguish “success” and “failure”, we need examples of both.

- But publication norms preclude exhaustive detail
  - No reports of “failure”
  - Only one report of “success” per compound
  - “Everything should work?”

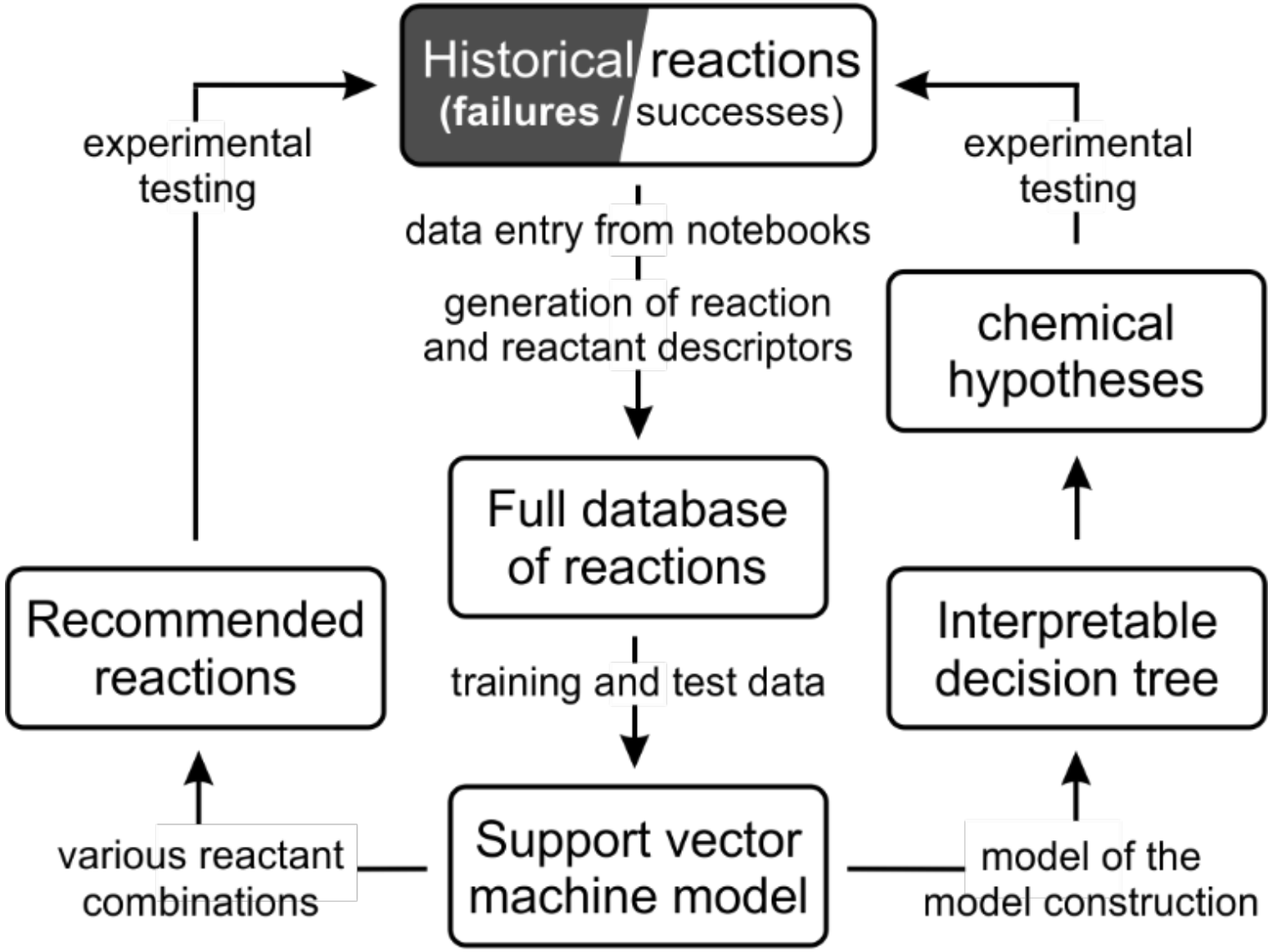
- The reactions <u>have</u> been performed and recorded…
  - … in stacks of old laboratory notebooks

- Dark Reactions Project
  - http://darkreactions.haverford.edu

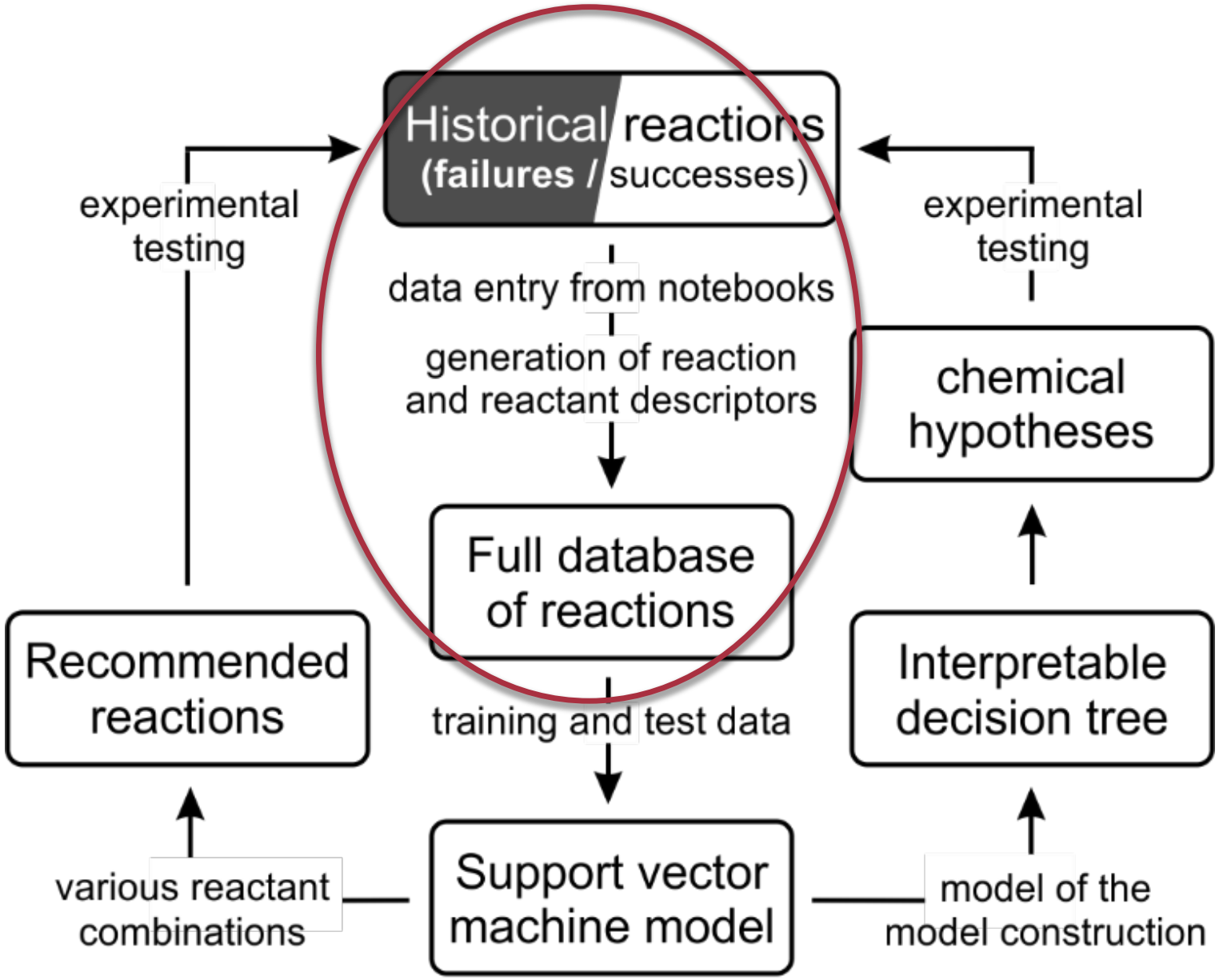




Nature **533**, 73-76 (2016)



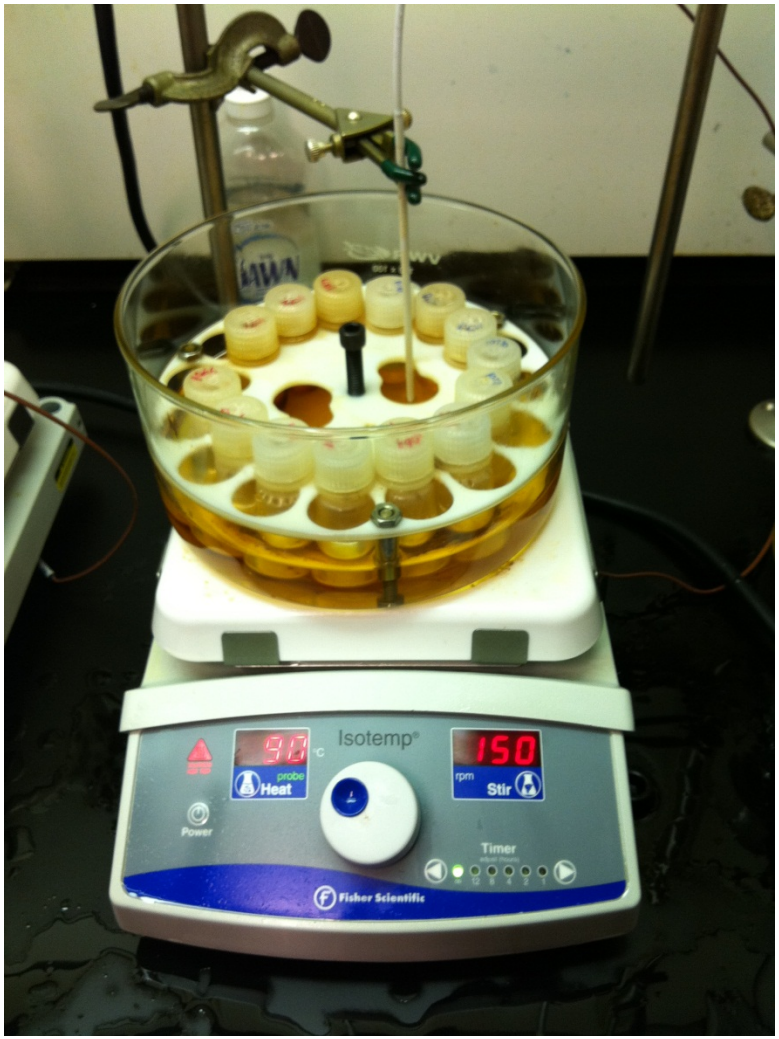
Nature **533**, 73-76 (2016)



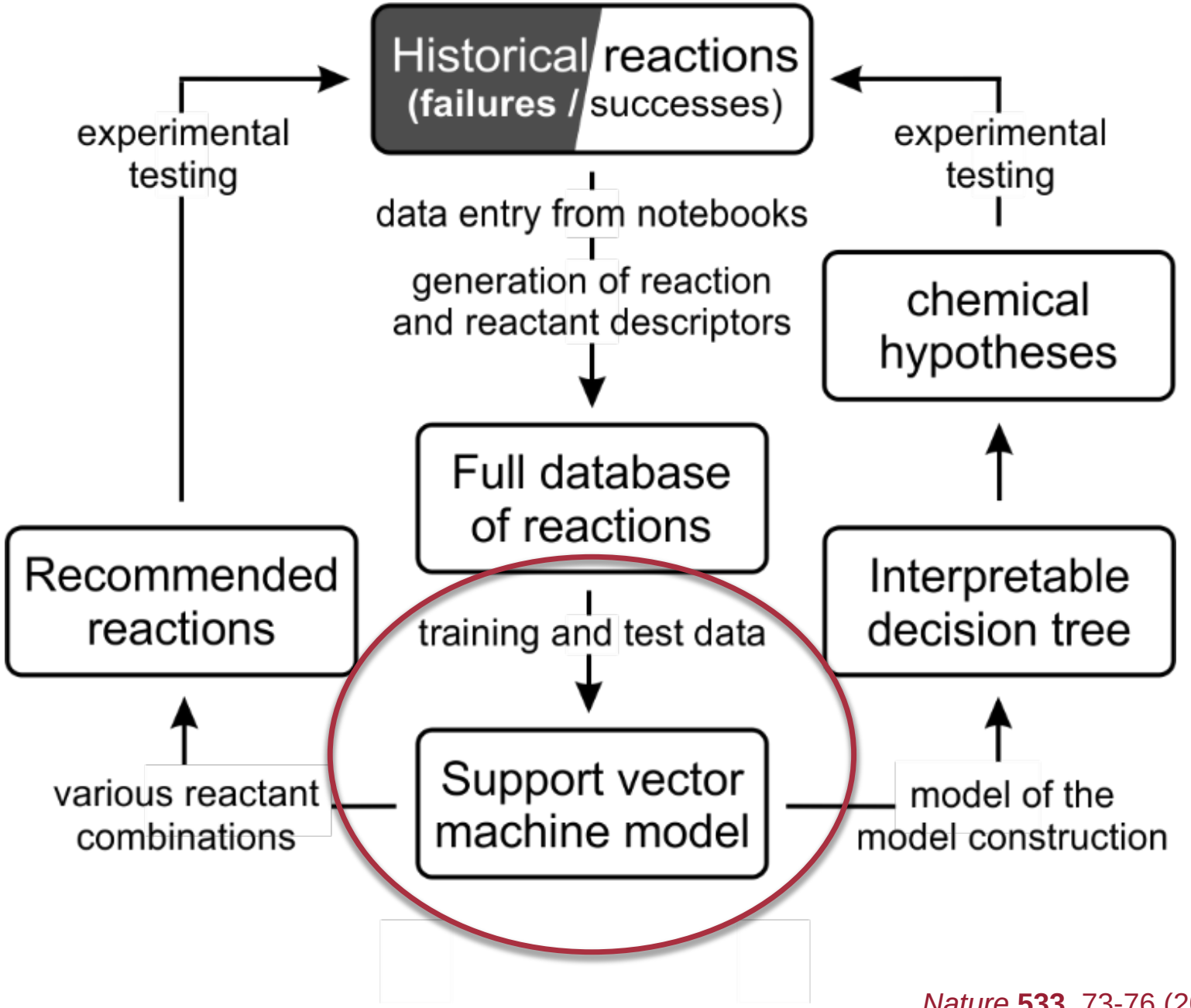
Nature **533**, 73-76 (2016)

Hydrothermal Synthesis

- Two metal sources
- One amine
- (Oxalic acid)
- Water
- pH
- Temperature
- Time

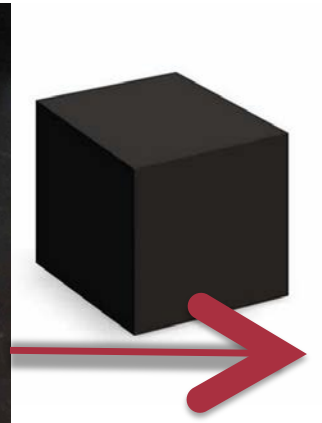


Simple database schema!

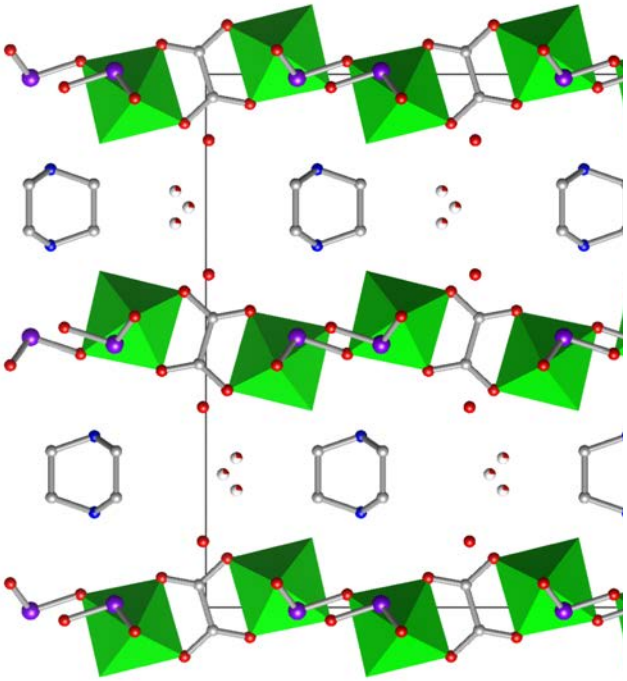


Nature **533**, 73-76 (2016)

You can't build a classifier on raw data...



"chemistry"

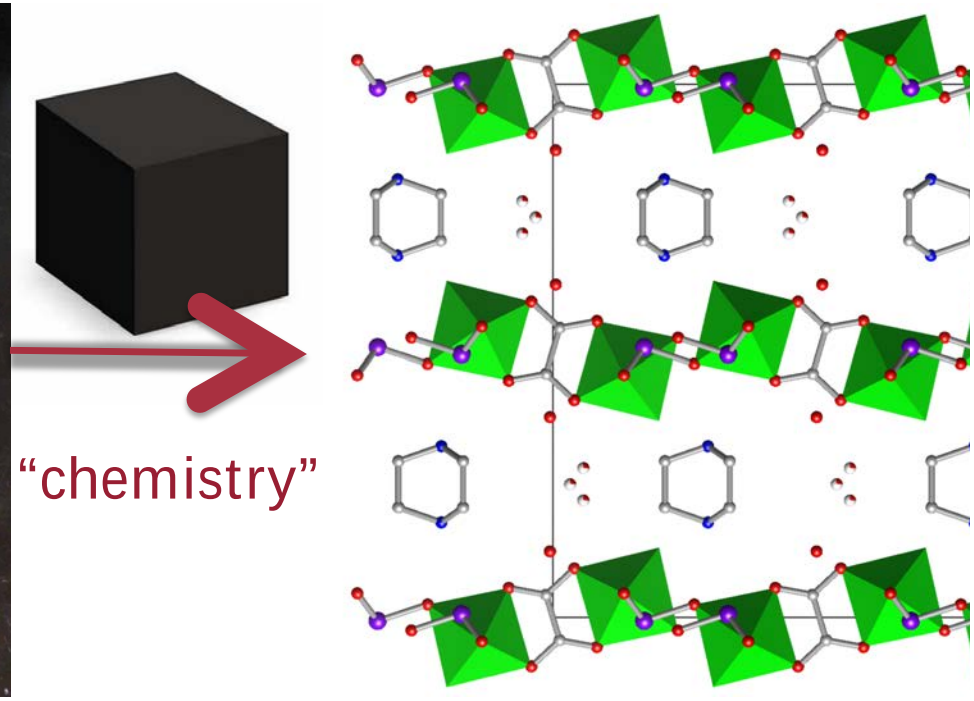


Inputs:
Concentration
pH
Time
Temperature
Etc.

"machine learning"

Outputs:
Big crystal
Small crystal
Something bad ("tar")
No reaction

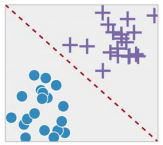
...you need an appropriate representation.



Inputs:
Concentratio
pH
Time
Temperature
Etc.

ChemAxon

327-dim
reaction
description
vector



Outputs:
Big crystal
Small crystal
Something bad ("tar")
No reaction

Turning lab data into descriptors

- Start with laboratory notebook data
  - Reagent names, masses, reaction conditions
  - Categorized crystal quality **outcomes** (1..4)
- Use cheminformatics to add physical properties
  - [Keep the reaction conditions, e.g., temperature, time]
  - Stoichiometries, ratios
  - Organic molecule properties
    - Solubility, polar surface area, # hbond donors/acceptors, etc.
  - Inorganic component properties
    - Electronegativity, ionization energy, radii, position in periodic table, etc.
- Use these as *reaction descriptors* (327 per reaction)

ChemAxon

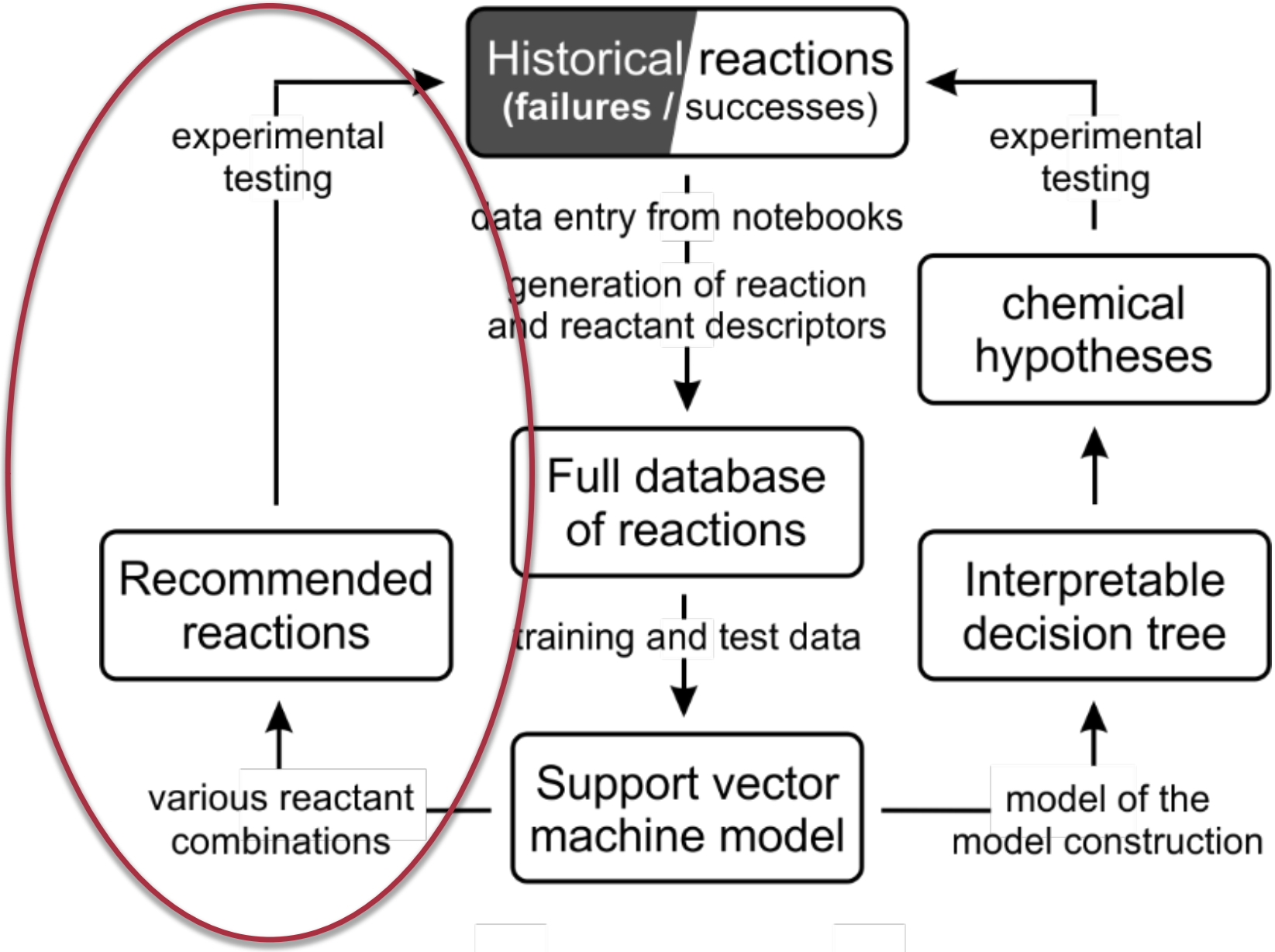
Nature **533**, 73-76 (2016)

Be careful how you test & train…

- Random test & train does <u>not</u> generalize!
- "Exploratory" split
  1. Put every set of reactions that share the exact same reactants into either the train or test set
  2. Tests how the model performs on "novel" reactions
     - Avoid "hidden" overfitting. See also relevant recent work by Max Hutchinson (Google) & Izhar Wallach (Atomwise)

•…and recent work in KDD '18 "Optimizing a Machine Learning System for Materials Discovery" (and GMN's thesis)

Nature **533**, 73-76 (2016)

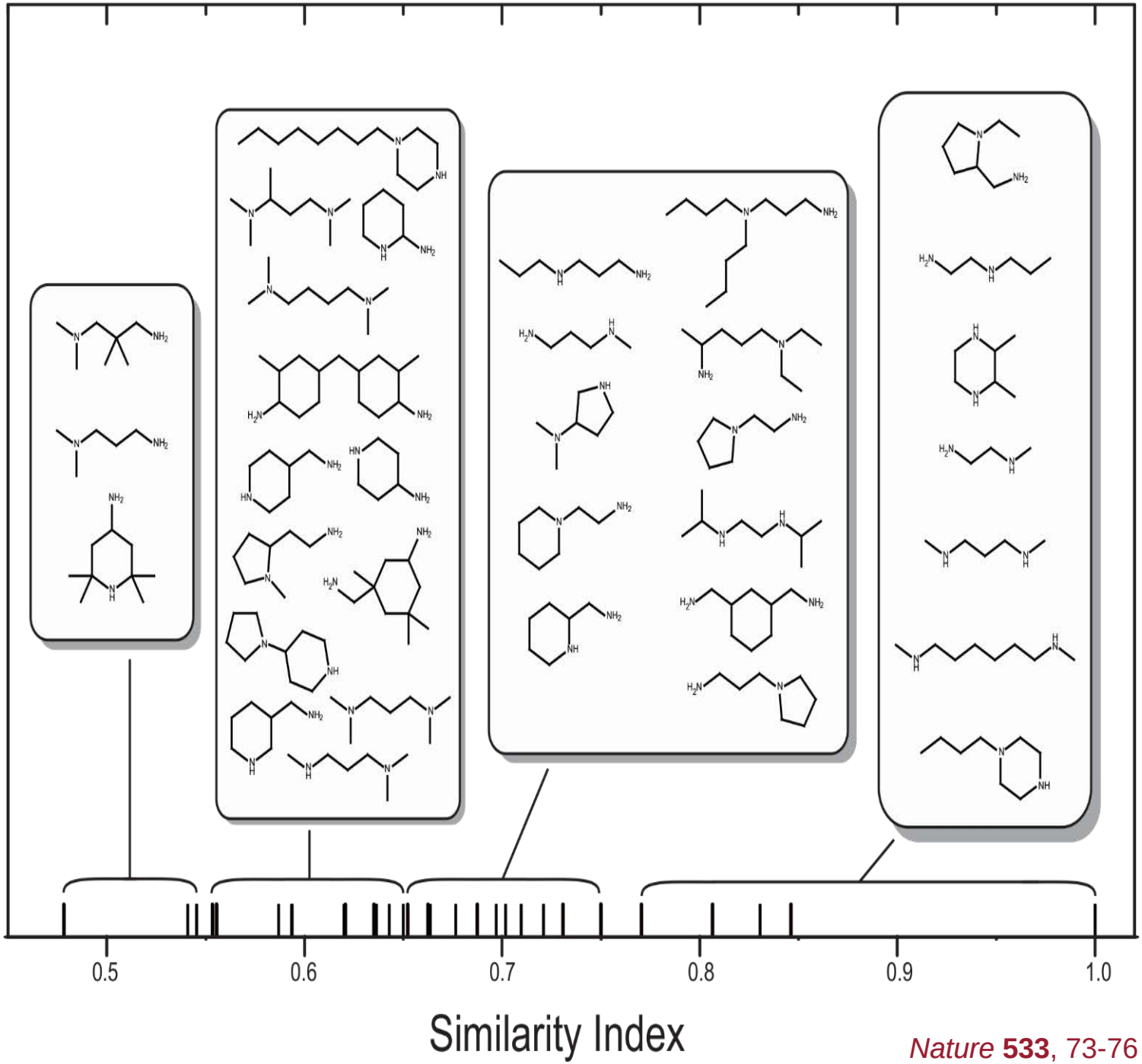


Nature **533**, 73-76 (2016)

Database-assisted Synthesis Planning

- Use databases of all commercially available organics
  - Emolecules.com
  - Keep only diamines, eliminate “bad” functional groups
  - Select 34/1681 based on price and availability
- Sample across structural similarity to existing amines
- Chemical novelty
  - “The chemistry of these 34 amines is essentially unknown in the formation of organically templated metal oxides, as indicated by the almost complete absence of such compounds in the Cambridge Structural Database (CSD). On average, 2 structures have been reported for each of the 34 amines, with 19 not existing in any templated metal oxide structure the CSD. In contrast, an average of 151 unique structures exist for the most frequently used amines (piperazine, ethylenediamine, 4,4’-dipyridyl, and dabco).”

Nature **533**, 73-76 (2016)



Nature **533**, 73-76 (2016)

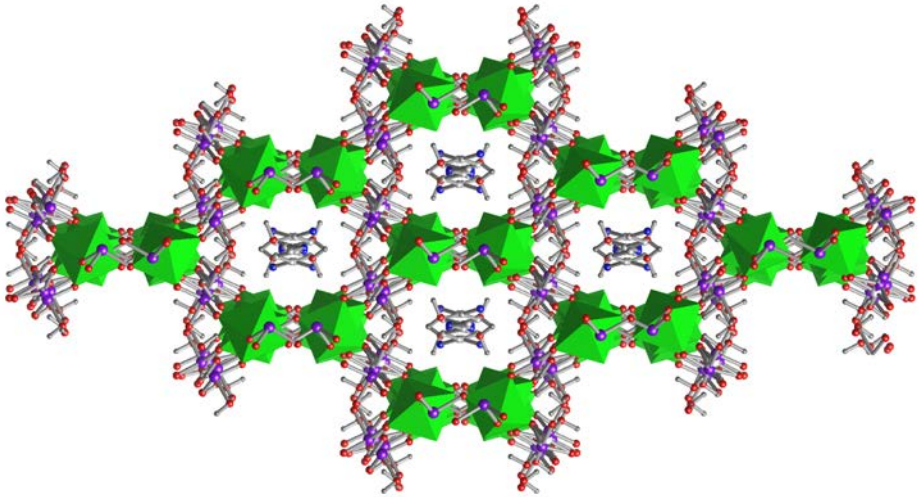
Human versus Machine

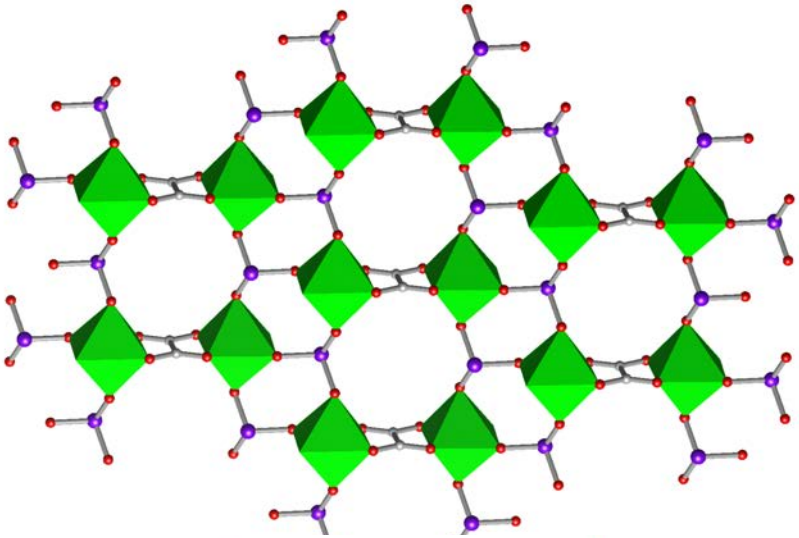
- Human expert's best "intuition":
  - Scale the mole quantities of the organic to match the successful "seed reaction"
- Model predictions
  - Scan through thousands of possible conditions
  - Computational black box (SVM) predicts outcomes
  - Do experiment on most promising outcome
- Tested with 495 new experiments

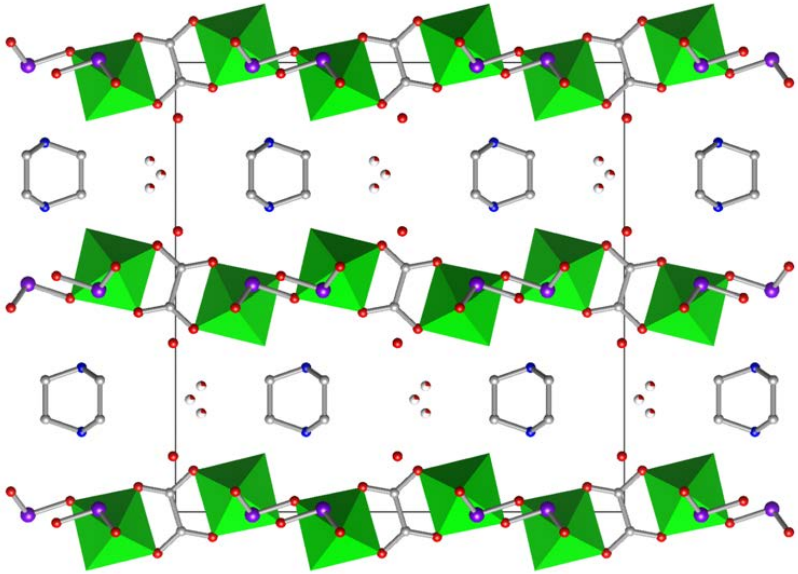
Katie Elbert '14 (UPenn)
Yunwen ('Helen') Yang '15
Wenjia ('Scott') Huang '15 (JP Morgan/Chase)
Aurellio Mollo '17 (Harvard)
Malia Wenny '17(Harvard)

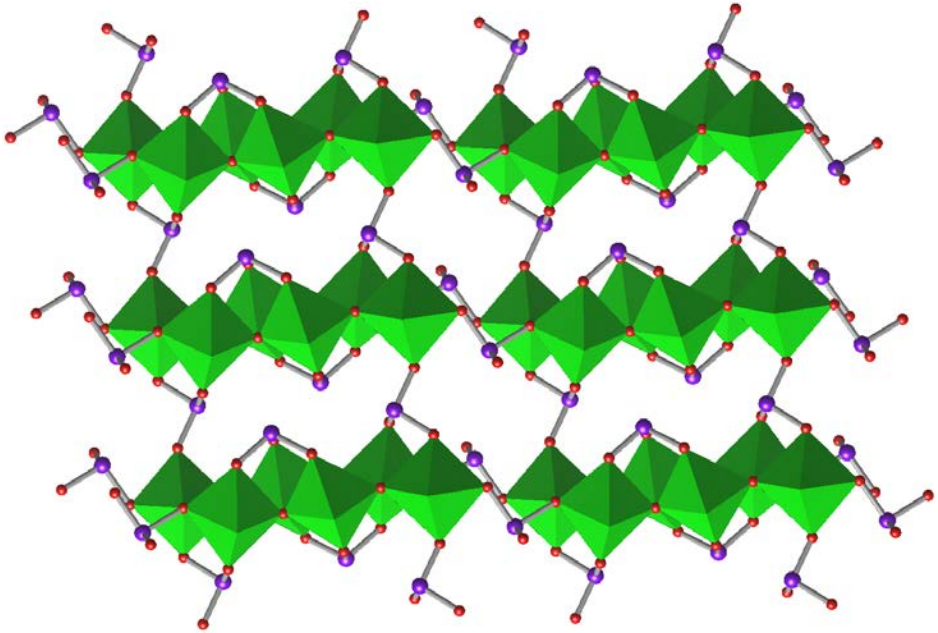
Nature **533**, 73-76 (2016)

We made many new compounds!



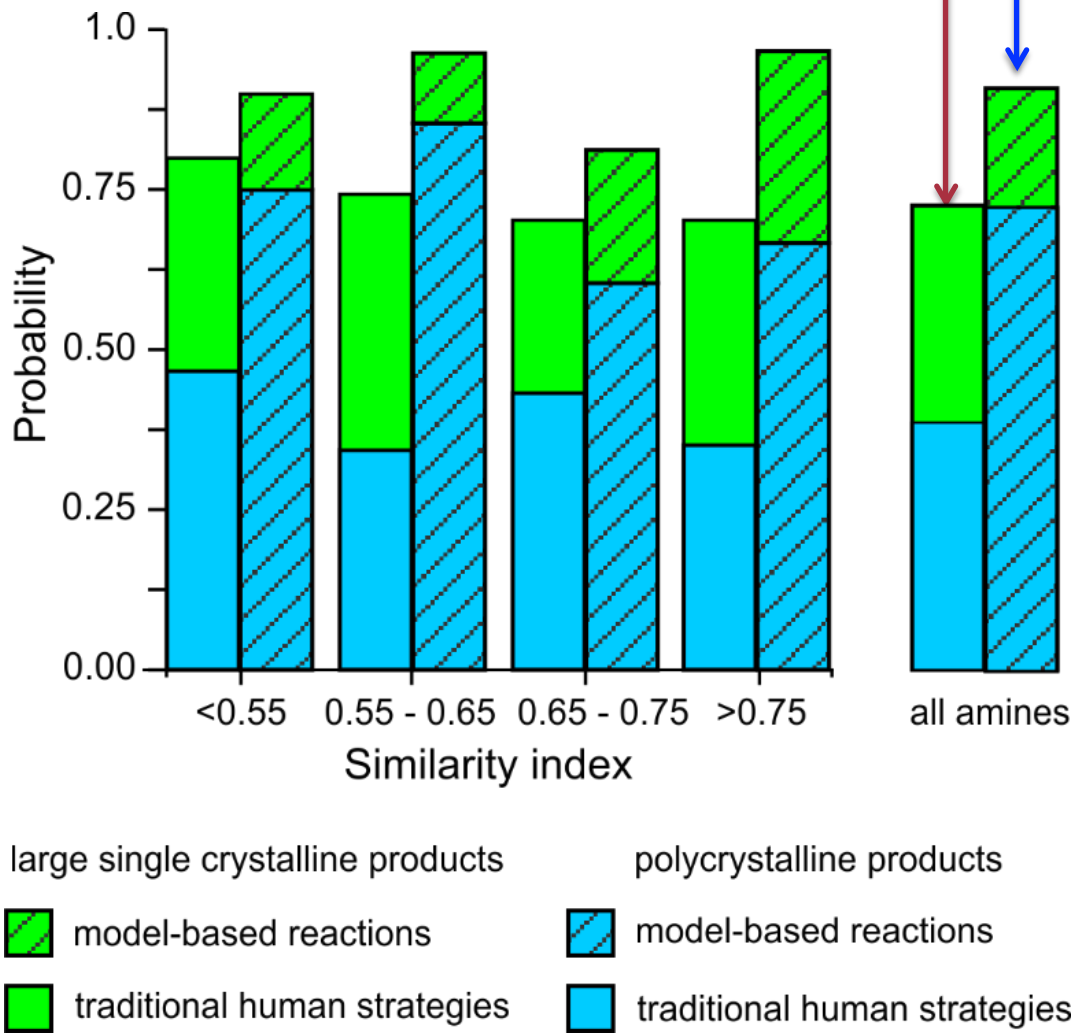






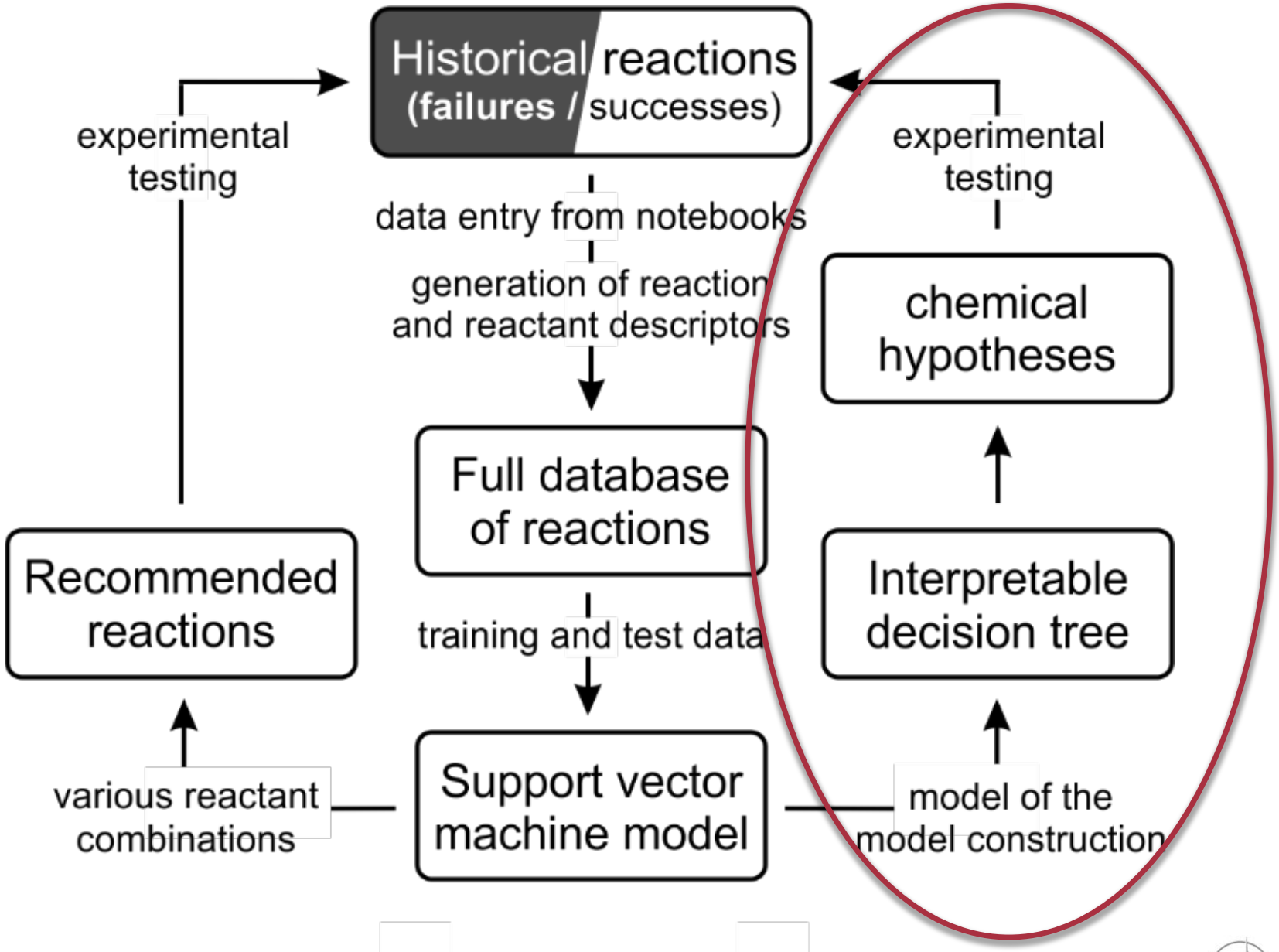
Human 78%

Machine 89%



Fischer exact test: $p<0.01$ that machine is <u>not</u> better than human

Interpretable ML models can reveal insight into reaction processes and provide new hypotheses—ML need not be a "black box".

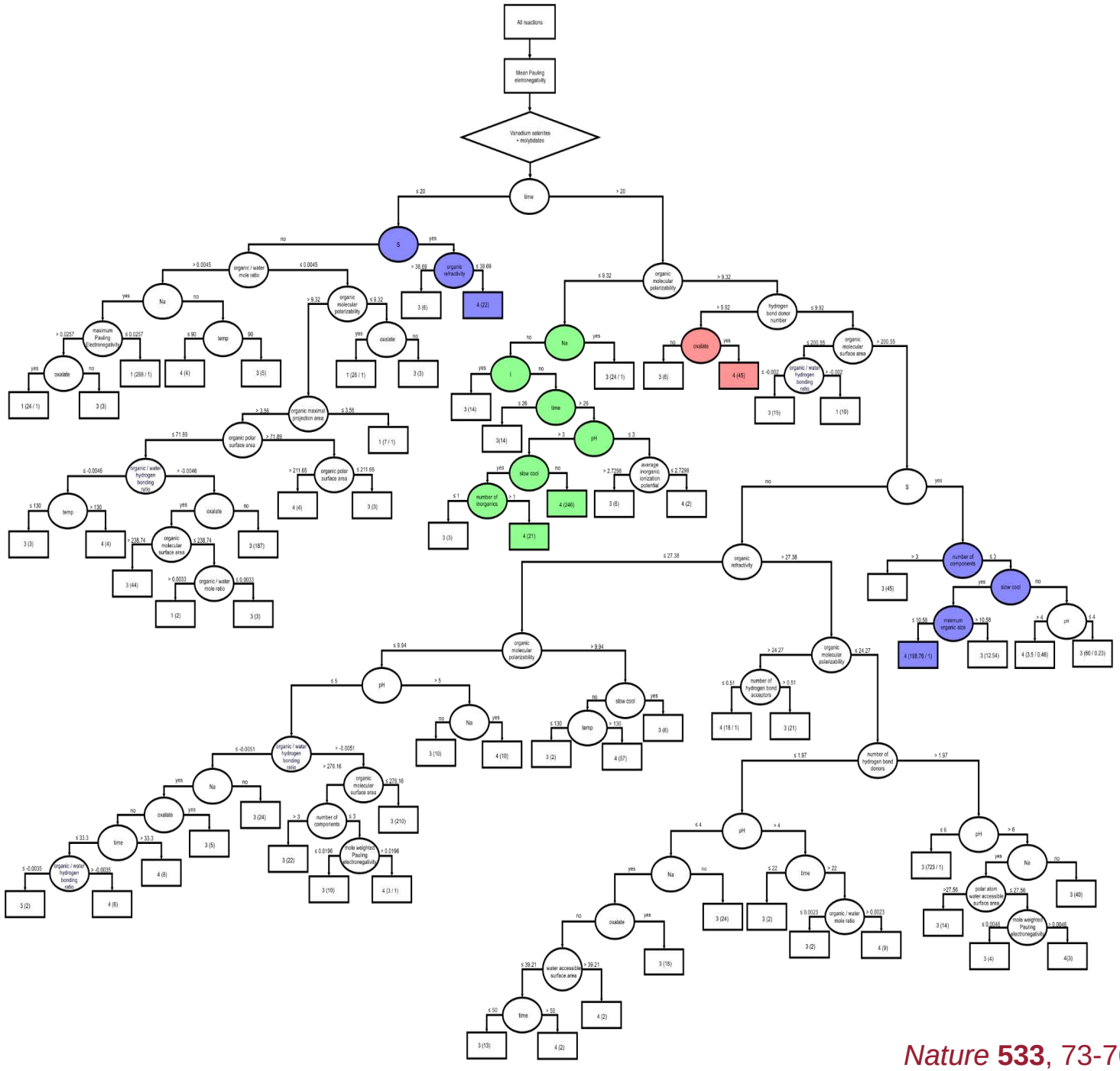


Nature **533**, 73-76 (2016)

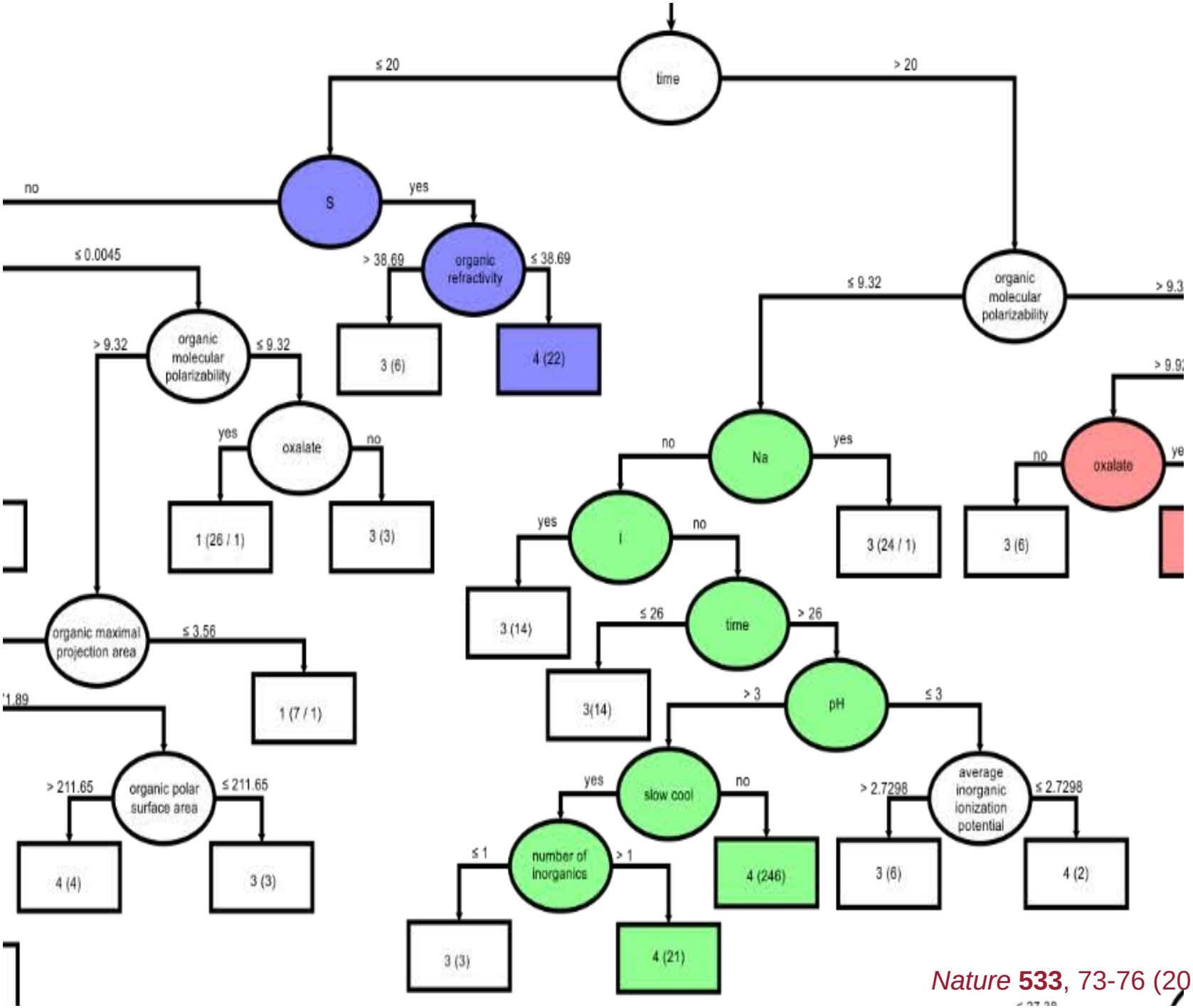
SVMs are not “human readable”

- SVMs lead to high-quality predictions
  - …but we can’t learn from them!
- Build a “model of the model”
  - Surrogate models
  - Train a decision tree that reproduces the SVM predictions

Nature **533**, 73-76 (2016)



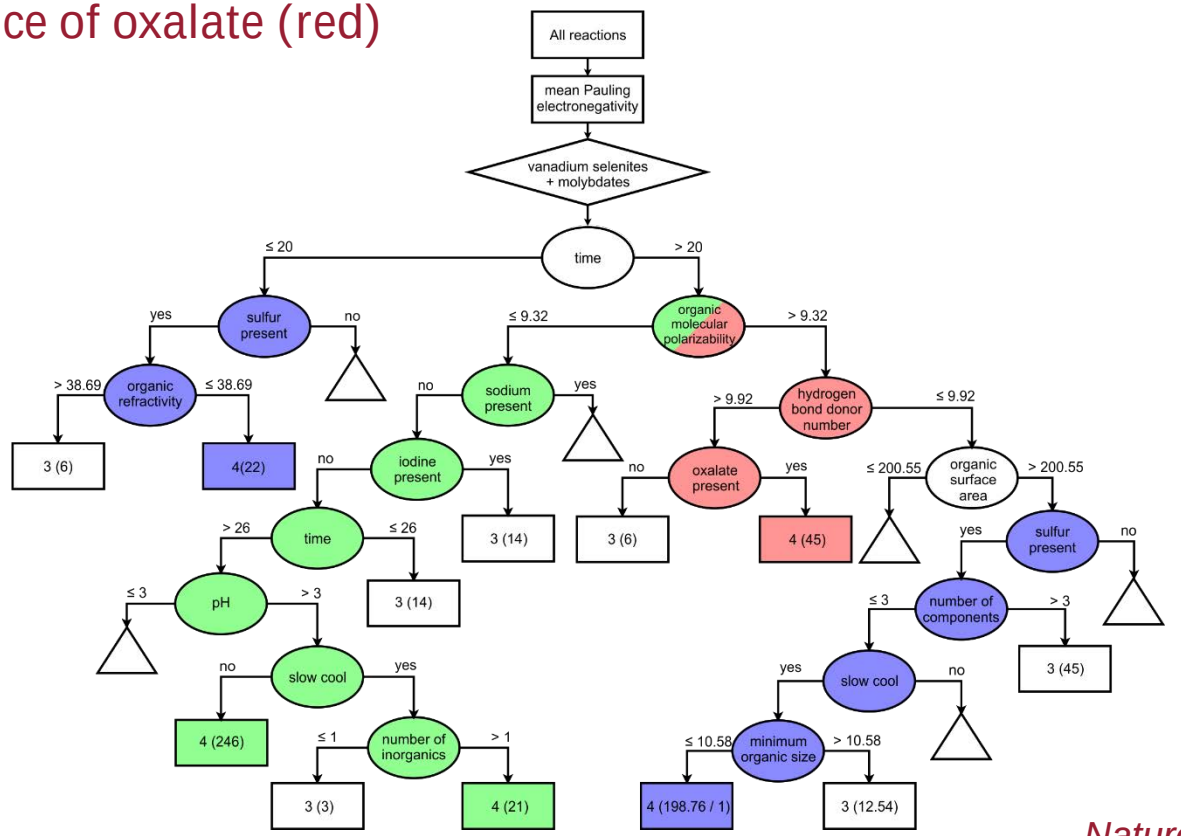
Nature **533**, 73-76 (2016)



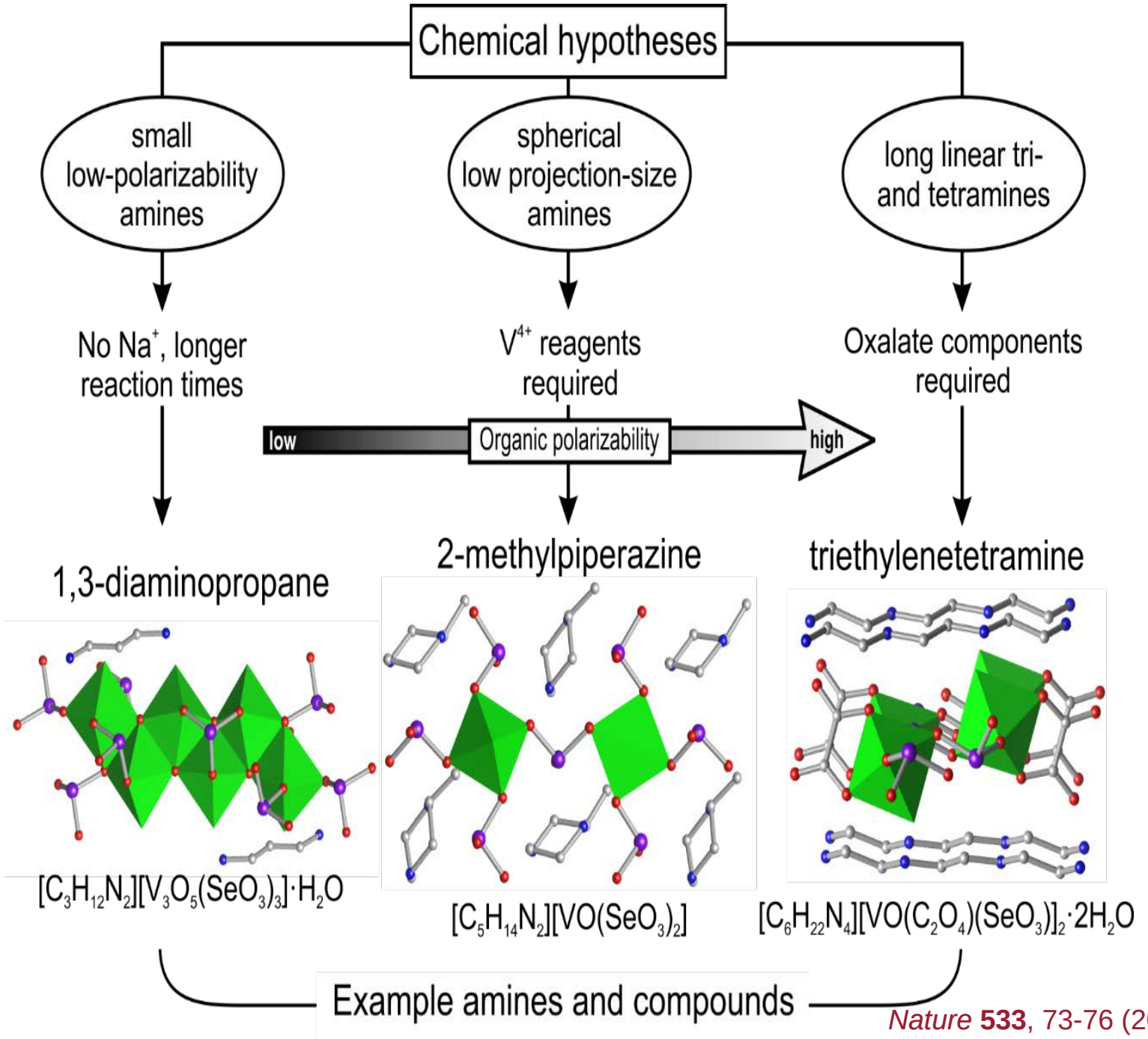
Nature **533**, 73-76 (2016)

What do the humans learn ?

- Low polarizability amines require no Na^+, longer reaction times, pH > 3 (green)
- Spherical, compact amines (with moderate polarizabilities) require the presence of S (blue)
- Long linear tri- and tetramines (with high polarizabilities) require the presence of oxalate (red)



Nature **533**, 73-76 (2016)

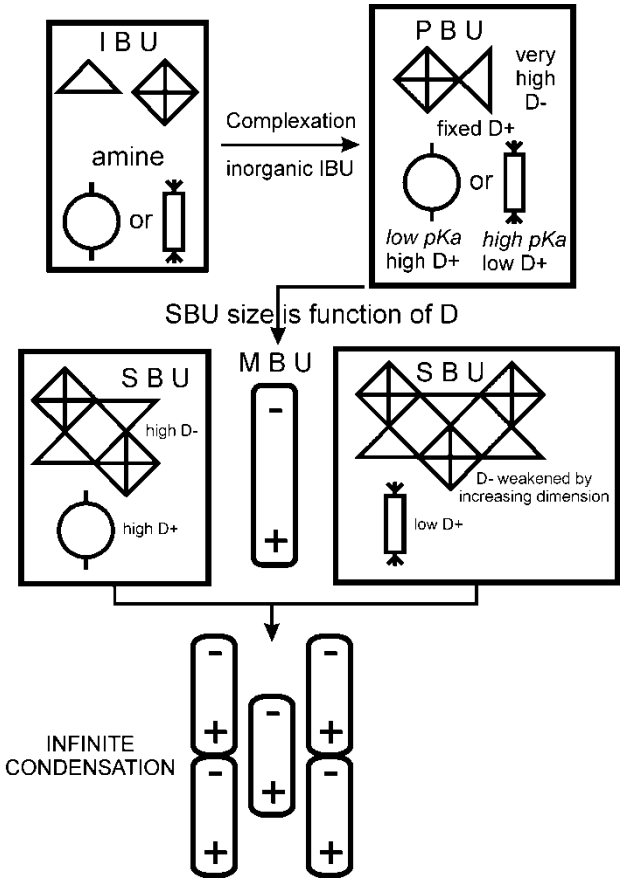


Nature **533**, 73-76 (2016)

ML can be used to generate ideas for new experimental plans, and optimal ways of verifying proposed hypotheses.

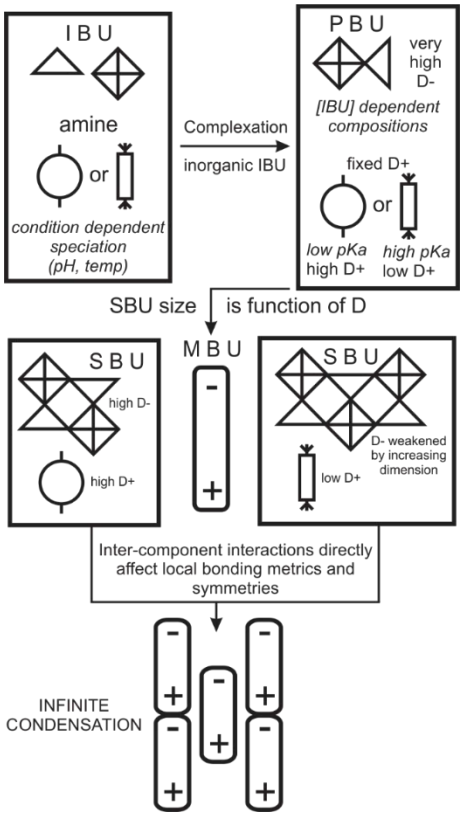
Can we test existing conceptual "models" of crystal formation?

- Primary: Reactant concentrations
  - Speciation and relative concentrations
- Secondary: Charge density matching
  - Fixed amine charge densities
  - Variable charge densities on the inorganic components
- Tertiary: Weaker influences
  - Sterics
  - Hydrogen-bonding
  - Symmetry



Ferey, G. *J. Fluorine Chem.* **1995,** *72*, 187.

Descriptors represent the theory…



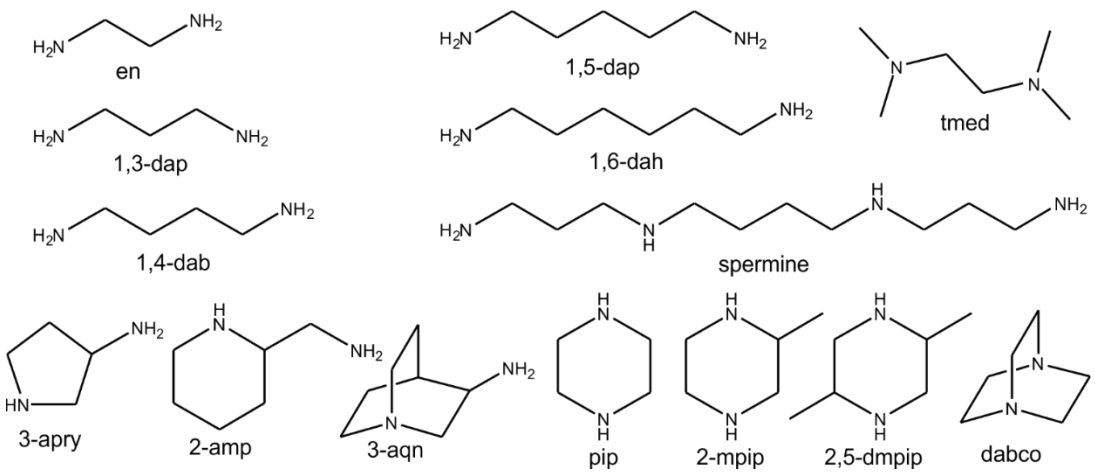
Adapted from Ferey, G. *J. Fluorine Chem.* **1995,** *72*, 187.

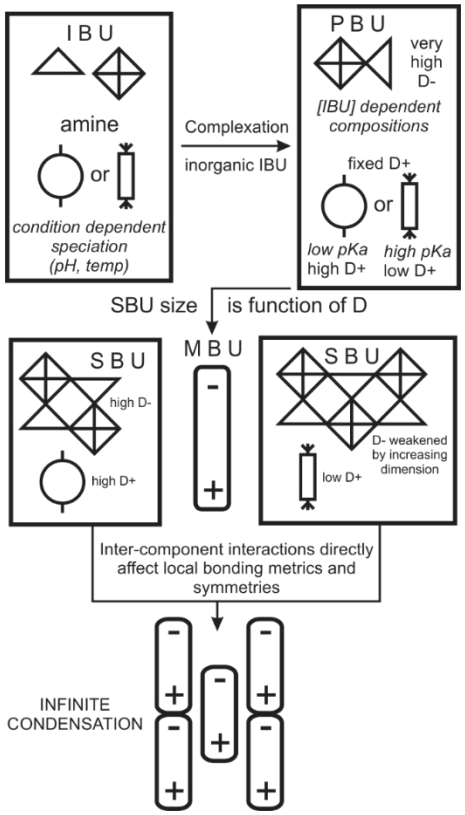
Type	Subset	Descriptor
Reaction	Stoichiometry	Amine amount (moles)
		Vanadium amount (moles)
		Selenium amount (moles)
	Conditions	Initial pH
		Time at maximum temperature
		Maximum temperature
Amine	Amine structure	Chain length
		Molecular weight
		Bond count
		Nitrogen count
		Primary ammonium site (yes/no)
		Cyclic structure (yes/no)
		Spherical (yes/no)
	Amine acidity	Minimum pK_a
		Maximum pK_a
	Charge density	Maximal projection area / nitrogen
	General properties	Reagent is an HCl salt (yes/no)
Inorganics	Vanadium counter ions	Vanadium source contains NH_4^+ (yes/no)
		Vanadium source contains Na^+ (yes/no)

Polyhedron **114**, 184-193 (2016)

...select amines to test the theory.

- 14 amines were selected using several criteria
  - Shape (linear, cyclic, bicyclic)
  - Amine properties (1°, 2°, 3°, mixed)
  - pK_a
  - Charge density

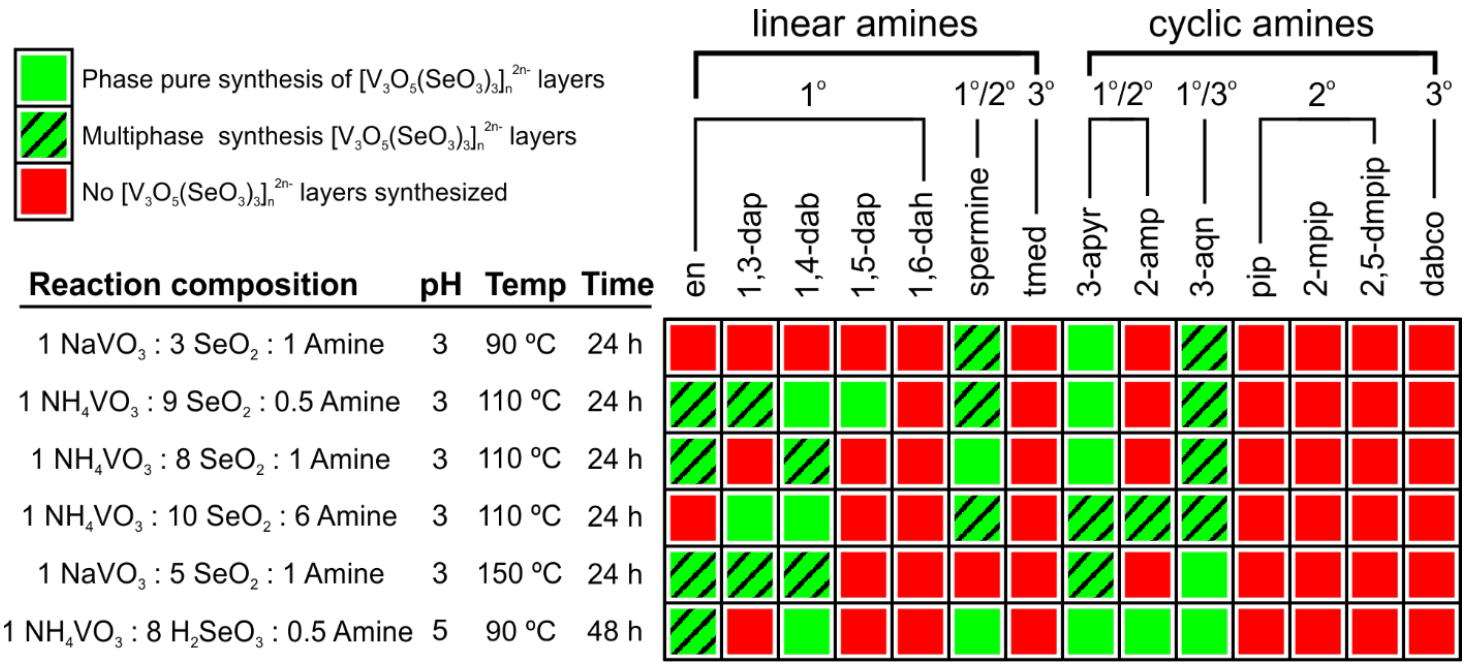




Adapted from Ferey, G. *J. Fluorine Chem.* **1995,** *72*, 187.

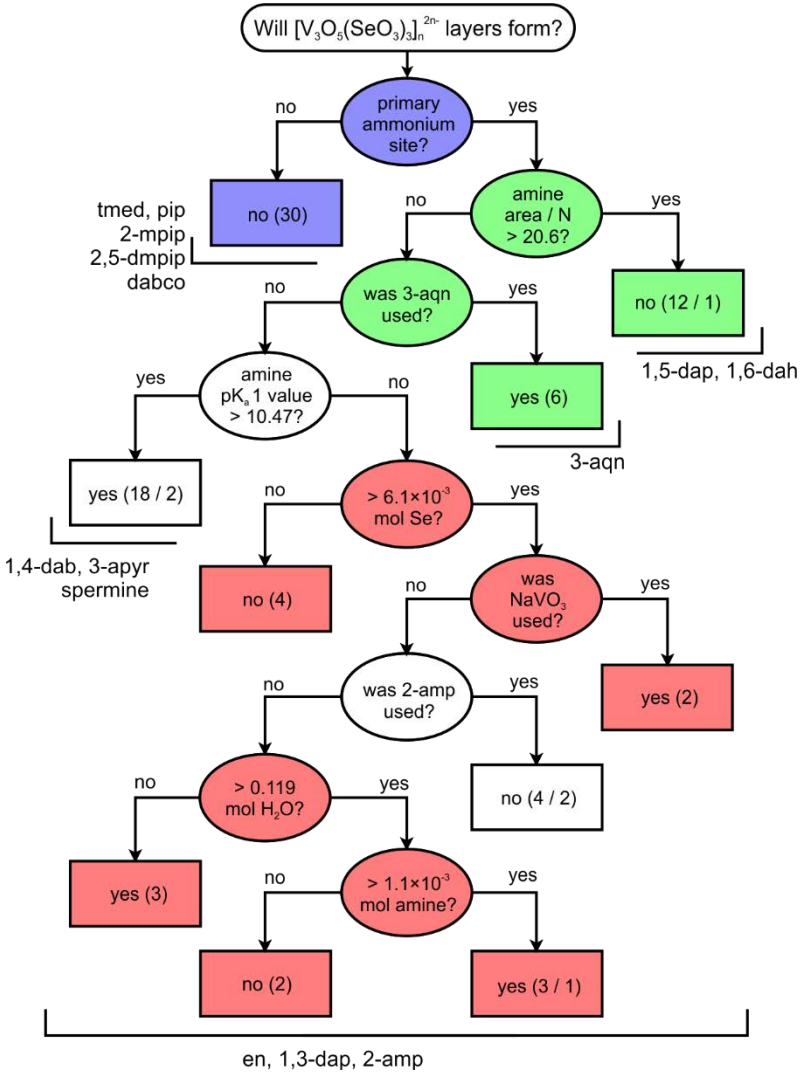
Polyhedron **114**, 184-193 (2016)

Send students into the lab to generate data...



Polyhedron **114**, 184-193 (2016)

Hierarchy of Influences = Decision Tree

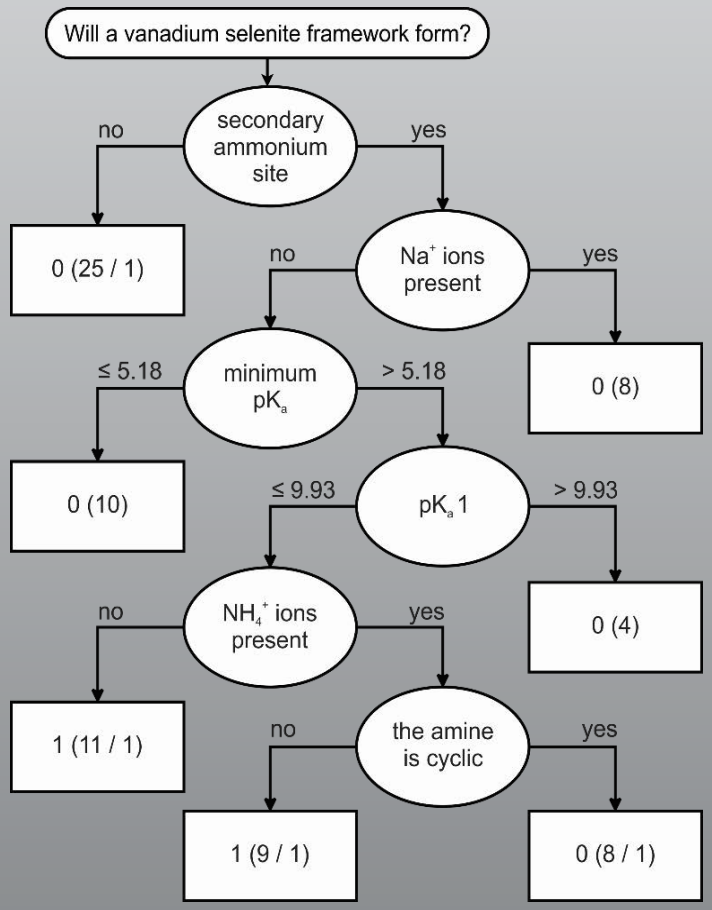


- C4.5 decision tree algorithm uses *information gain* as the criteria for selecting branches
- Use this to generate the *hierarchy of influences* supported by the experimental data

Polyhedron **114**, 184-193 (2016)

What amine properties govern 2D-layer formations?

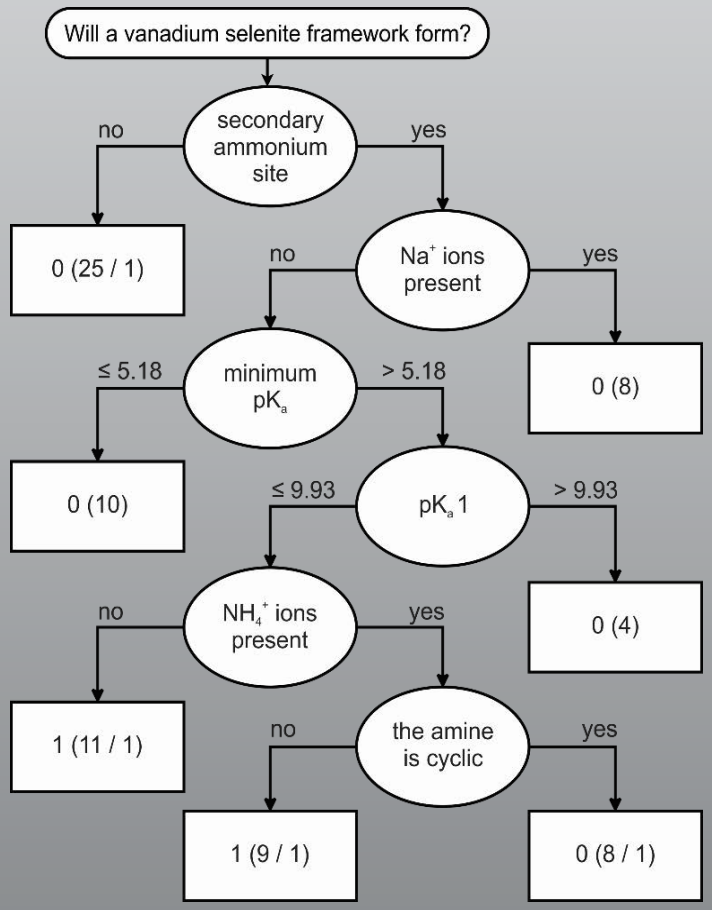
Historical dataset: 75 experiments



Molecular Systems Design & Eng. (2018) DOI: 10.1039/c7me00127d

What amine properties govern 2D-layer formations?

Historical dataset: 75 experiments

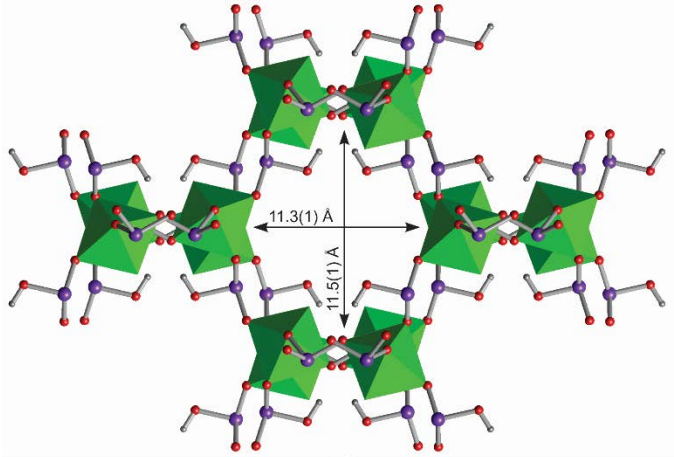


Use these reactant properties
+
stoichiometry variations

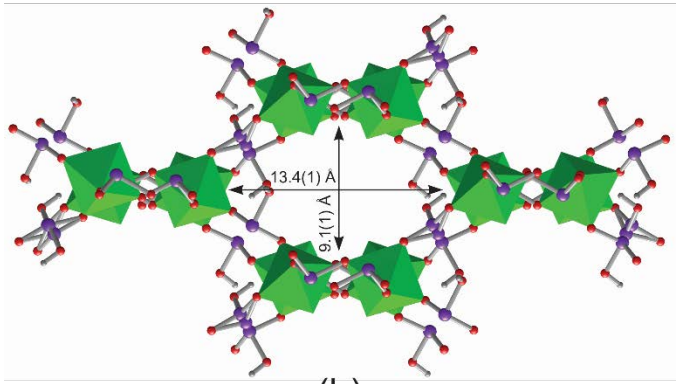
...to design a fractional factorial design experiment (128 expts) (23 amines) that *creates* the dataset we would need to build a model

Molecular Systems Design & Eng. (2018) DOI: 10.1039/c7me00127d

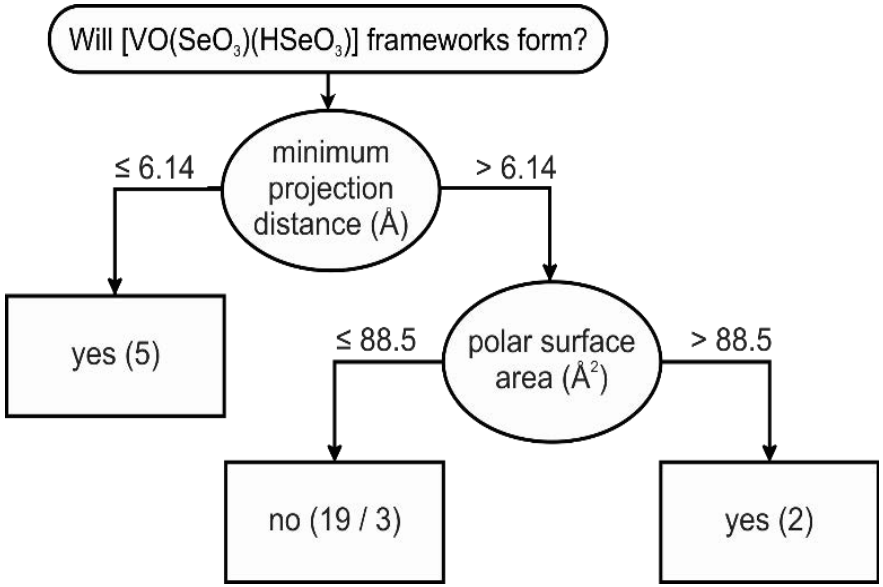
Two new layers types appear…



(a)

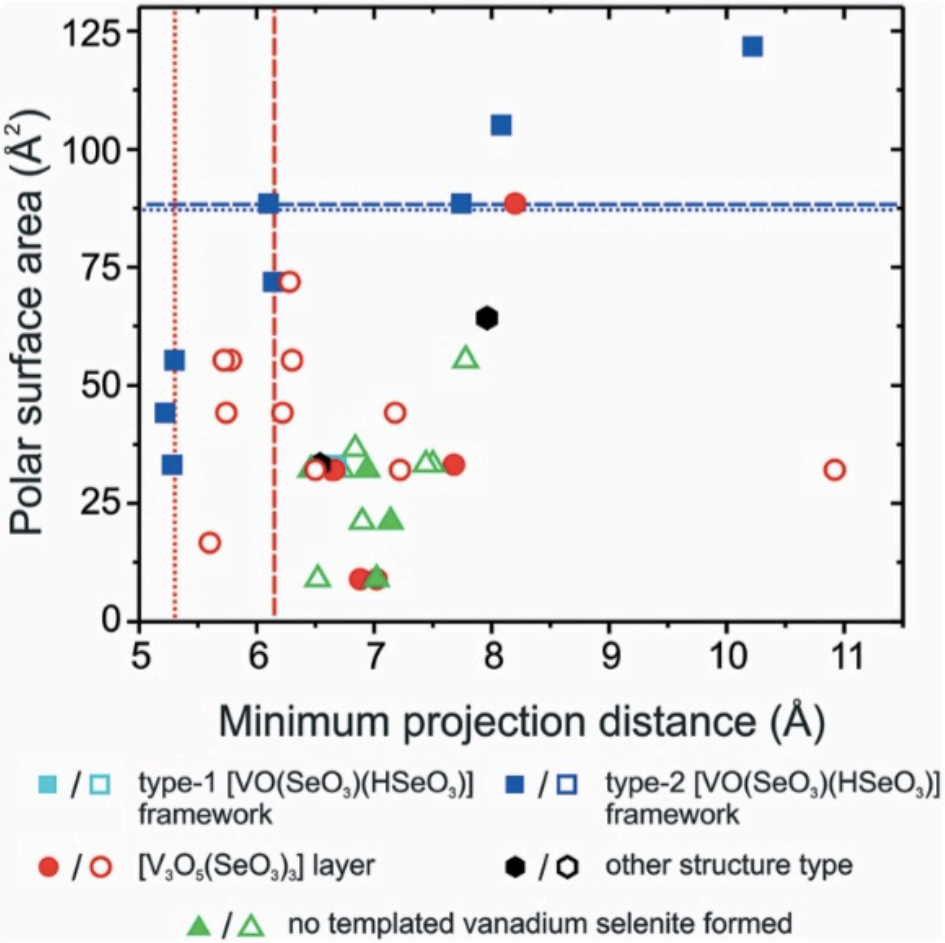


(b)



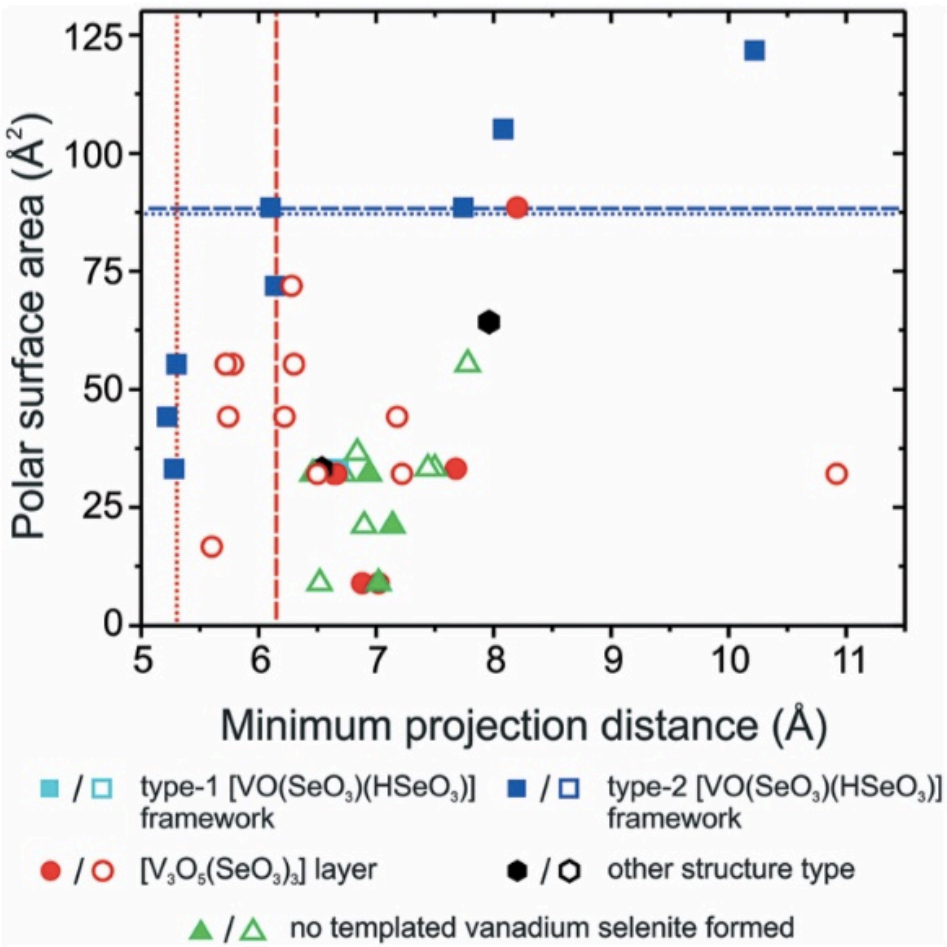
Molecular Systems Design & Eng. (2018) DOI: 10.1039/c7me00127d

What reactants best test the model?



Molecular Systems Design & Eng. (2018) DOI: 10.1039/c7me00127d

What reactants best test the model?



Identify 19 new amines near the decision boundaries

Run the experiments!

Refine the decision boundary

Molecular Systems Design & Eng. (2018) DOI: 10.1039/c7me00127d

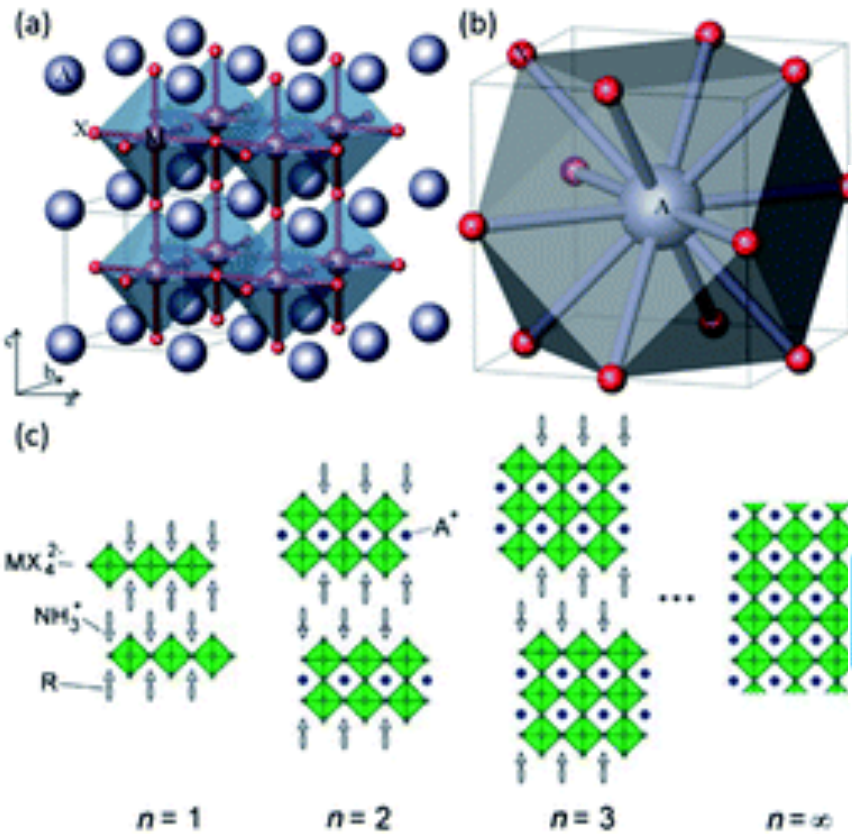
A dream
(and work in progress)

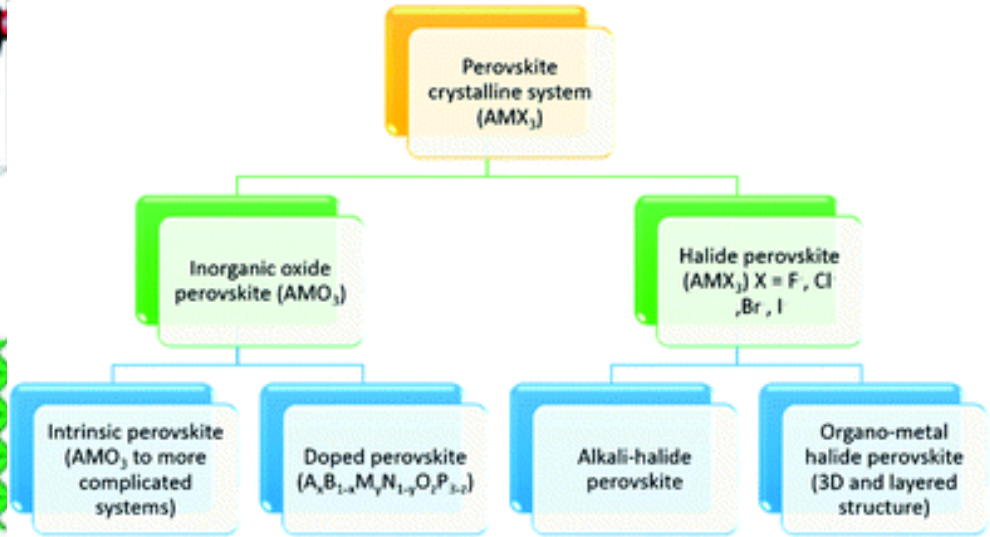
"Self-driving Chemistry" Semi-autonomous materials discovery

- Robotic synthesis, 10^3 experiments per day…
  - API for performing experiments and storing/retrieving results (including experimental metadata)
- …driven by Active Learning + Interpretable Models
  - Iteratively choose "optimal" experiments at the exploration/ exploitation frontier, based on model uncertainty
  - Get human input/veto on proposed experiments
    - Quantify "value added" of human experts
  - Output human readable "rules"
- …for a system with unclear design space
  - e.g., Organohalide perovskites
    - faster experiments
    - tremendous structural diversity (by varying organic components)
    - photovoltaics/sensors

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

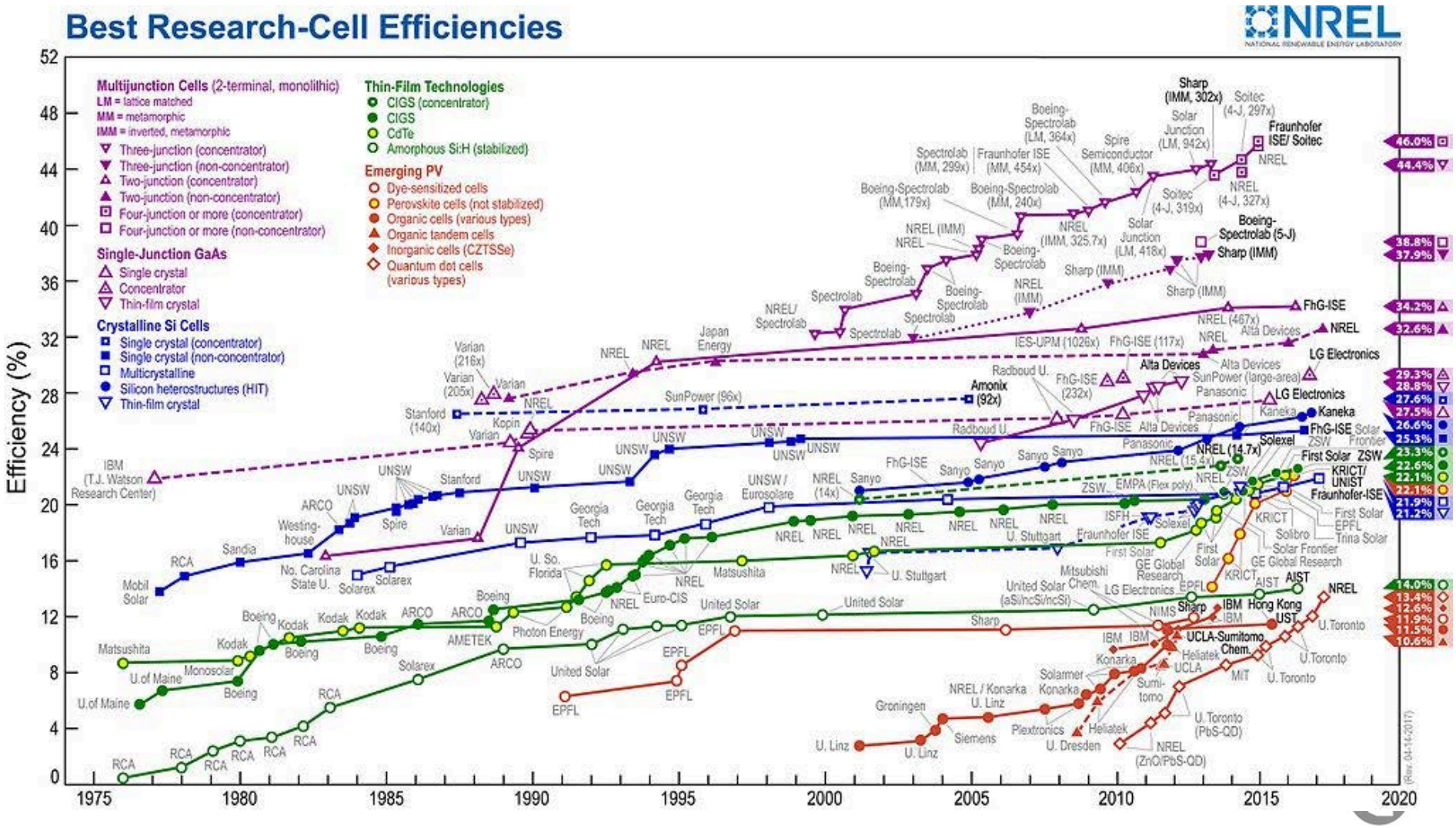
Organohalide hybrid perovskites are an emerging class solution-processable semiconductor of materials...



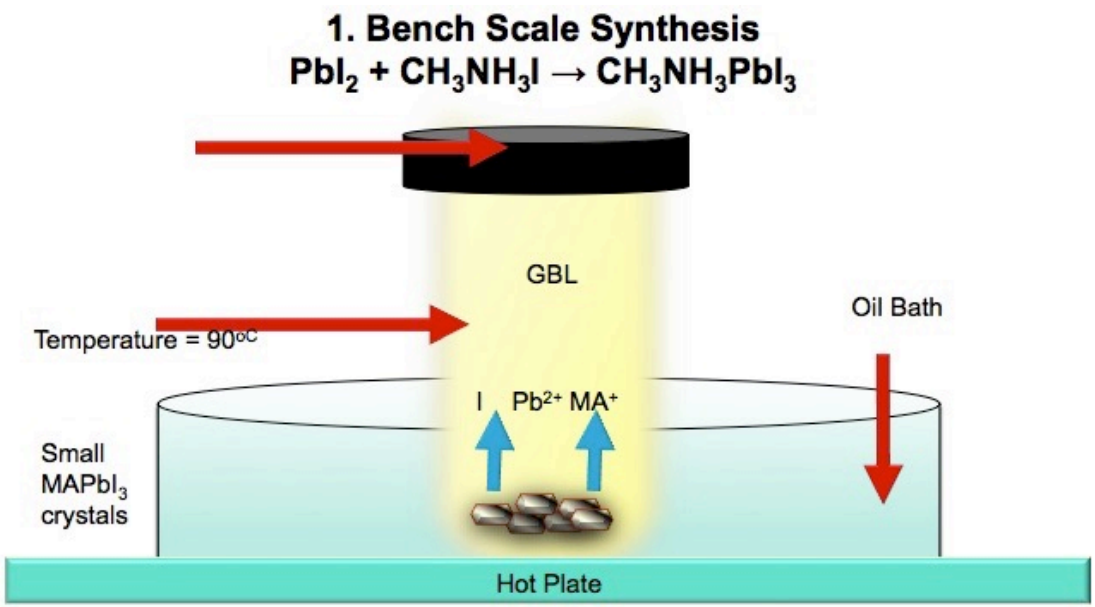


Source: DOI: 10.1039/C4EE00942H (Review Article) *Energy Environ. Sci.*, 2014, 7, 2448-2463

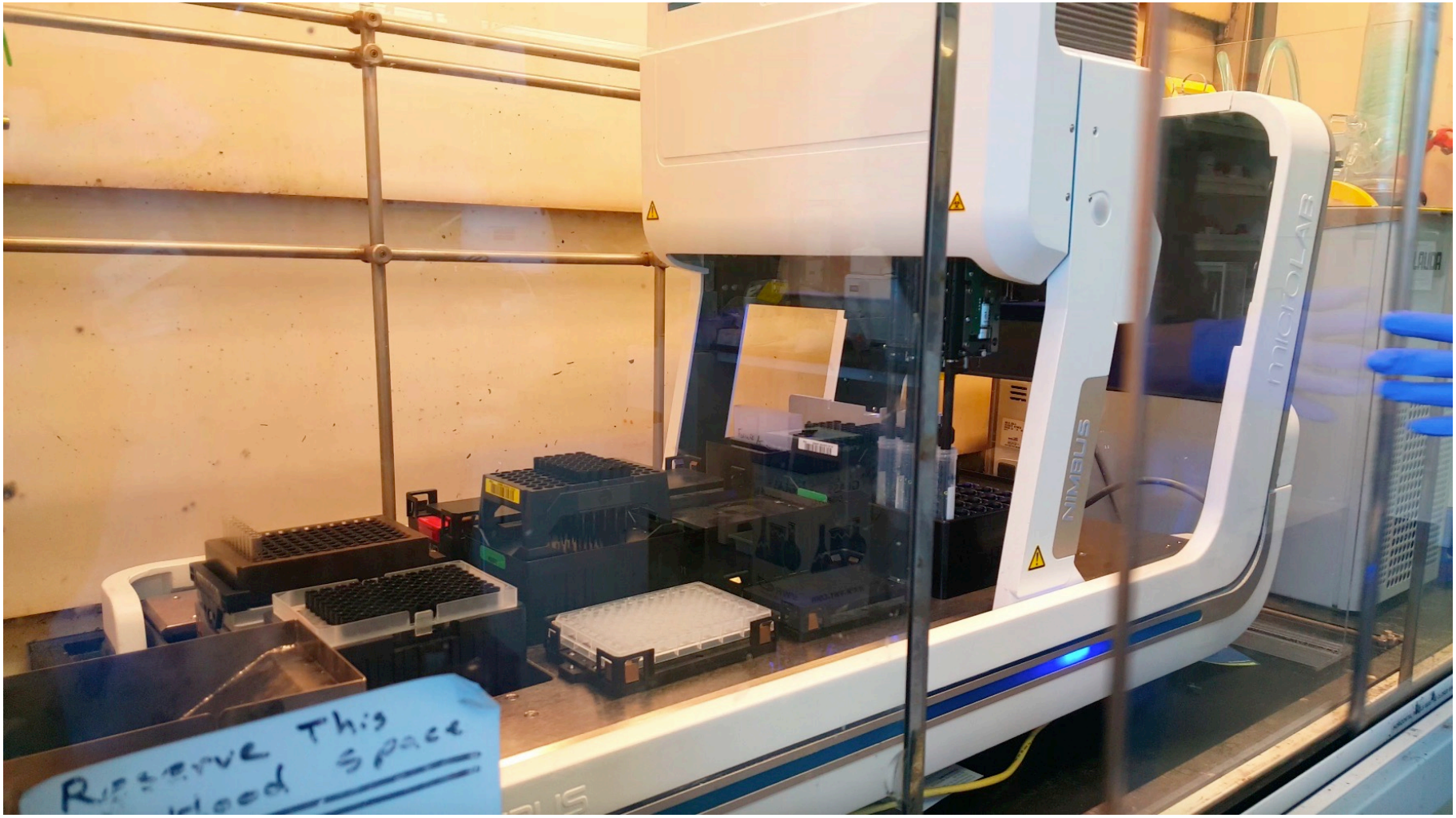
...with promising applications to photovoltaics, etc.



“Ordinary” perovskite synthesis



Our approach: Robots



Liana Alves ’18
Alyssa Sherman ‘18
Peter Cruz Parilla ‘19

“Robot ready” perovskite synthesis

Method
- Strong acids damage robot
- Reaction choice: Inverse Temperature Control and Antisolvent

Solvent
- DMF temperature too high for heating apparatus and crystal growth too long
- Gamma-butyrolactone (GBL)

Stock
- PbI2 insoluble in GBL
- MAI:GBL Stock Solution >1.2 M

Heating
- Inconsistent heating block impacts single crystal growth
- Temperature gradient 80-100 C

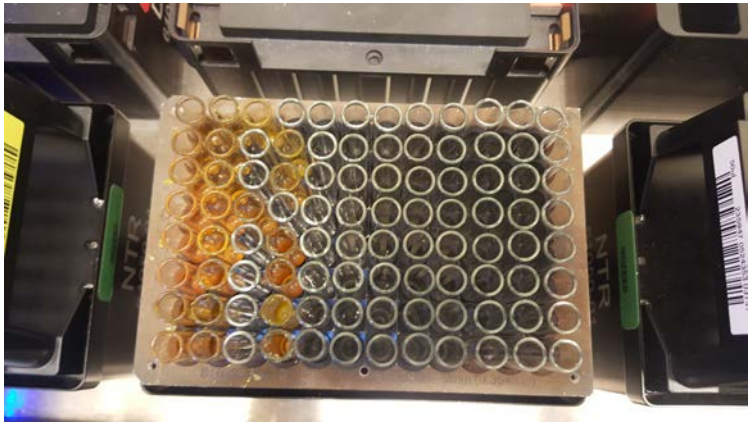
FAH
- Formic Acid (FAH) % vol needed and prevent PbI2 precipitate
- Shake 15 mins < 12.5 % vol of FAH

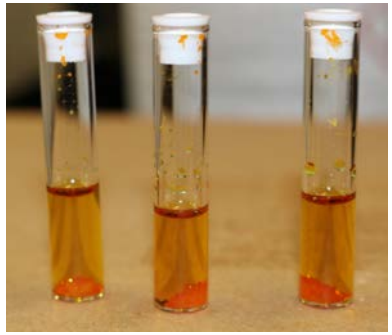
Filter
- How to filter 96 plate before dissolution ITC
- Vacuum filtration plate Glass 1.0 μm

Analysis
- Move crystals to plate for characterization
- Blow air into vacuum manifold, scotch tape XRD, machine a filter paper apparatus

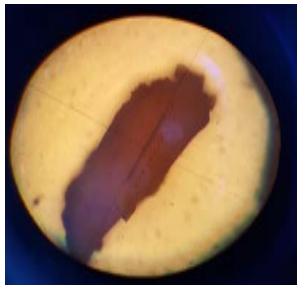
- DOE VFP @ LBNL
  - **Emory Chan (Molecular Foundry)**
  - Liana Alves ’18

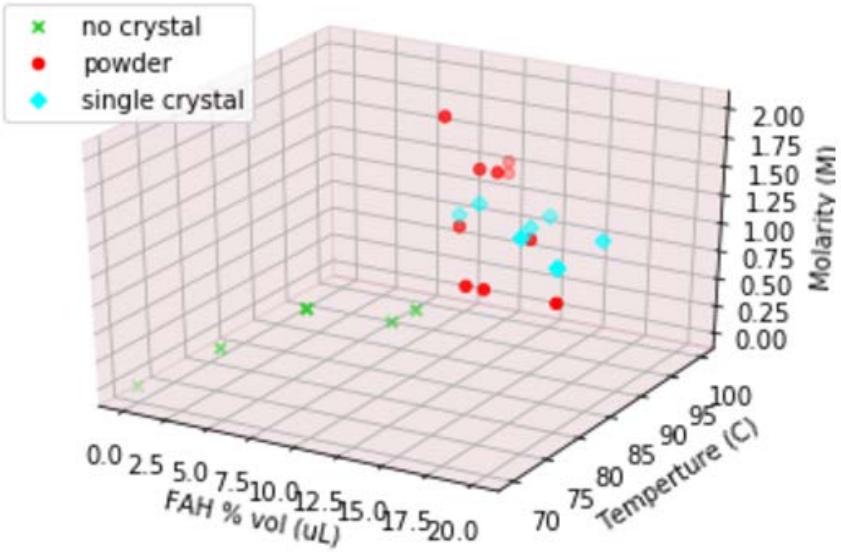
Sample data: Crystal formation





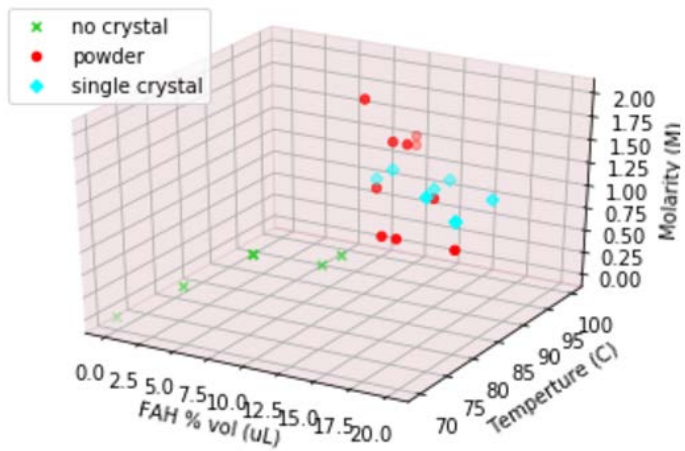
Optical microscopy

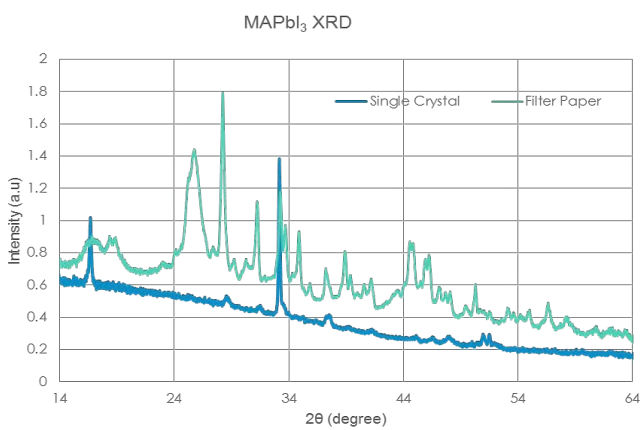


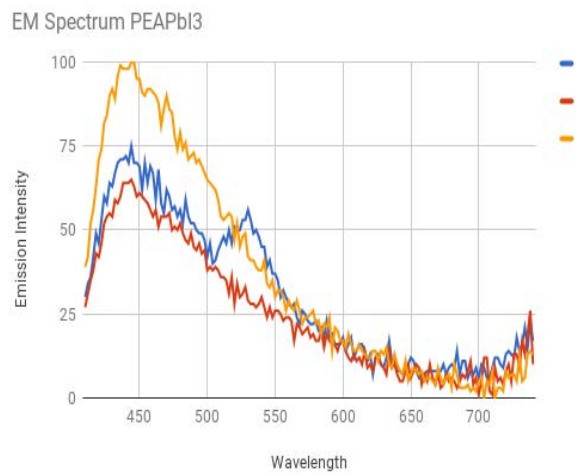


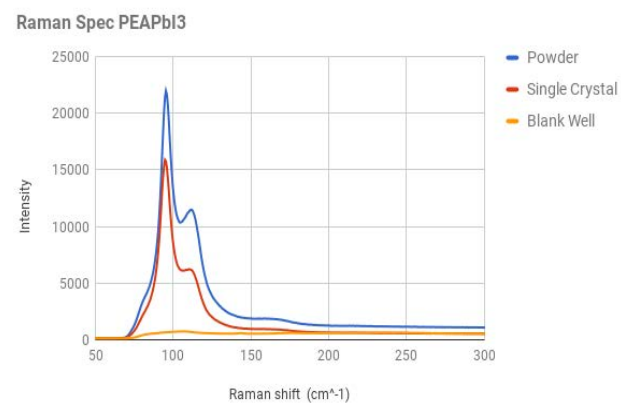
Liana Alves ’18 (→UCSD)
Peter Cruz Parilla ‘20
Alyssa Sherman ’18 (→UPenn)
Emily Brown ‘19

Sample data: Crystal quality & properties



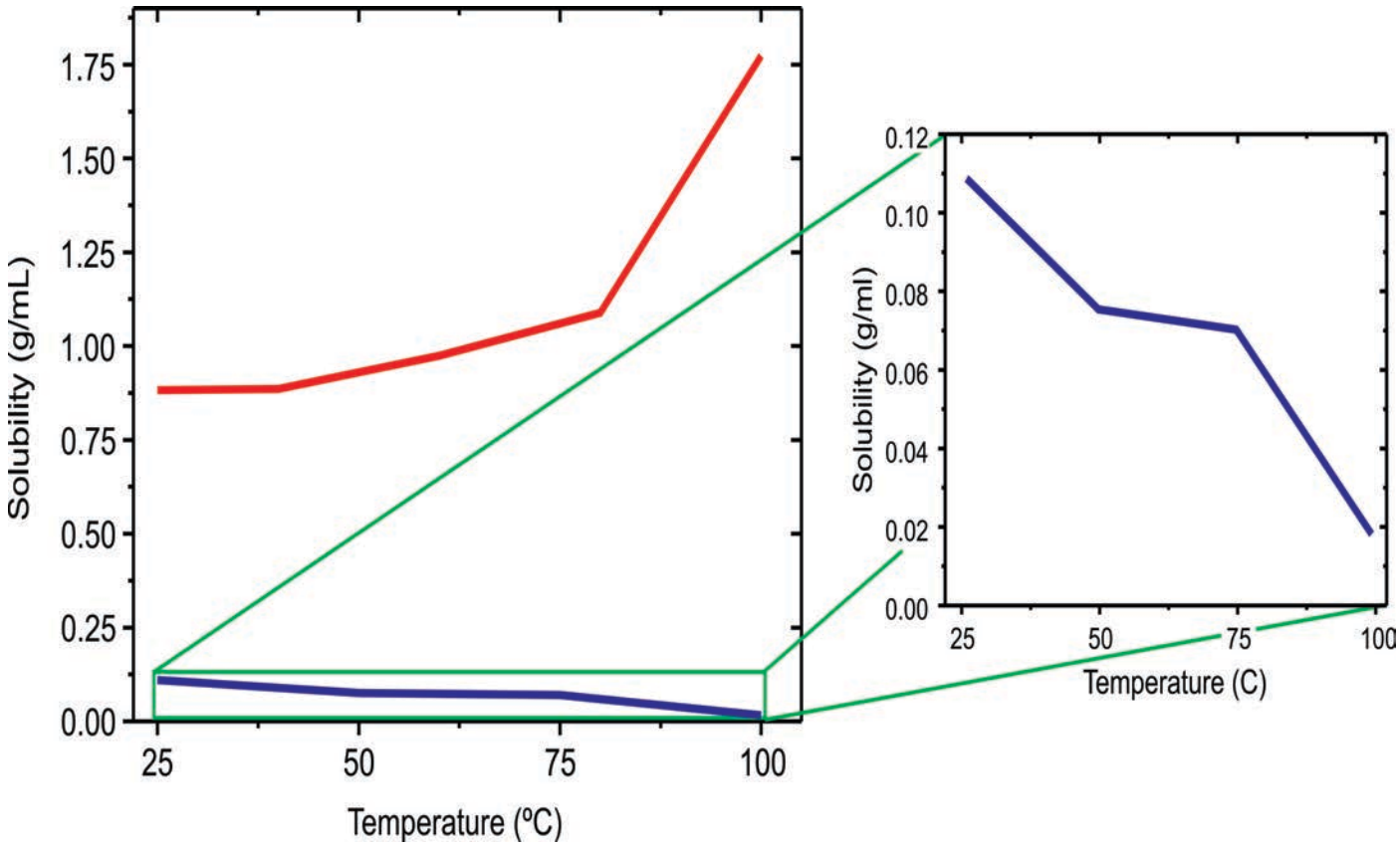






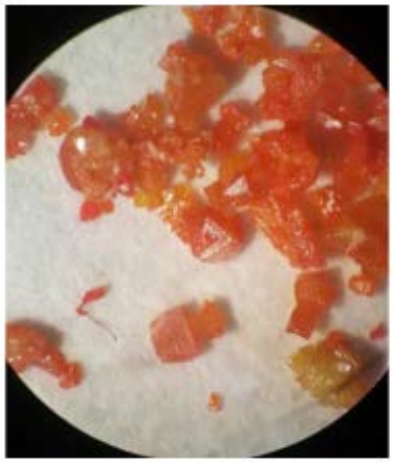
~800 rxns in 4 days—discover new crystal phase…story in progress…

Early finding:
New retrograde soluble phase of phenethylammonium lead iodide



B

A



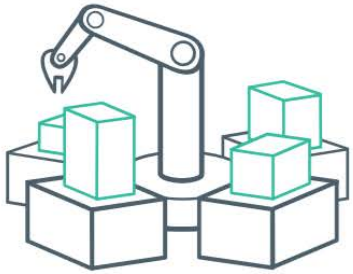
© 2017 Emerald Cloud Lab. All rights reserved.

Emerald Cloud Lab Overview

The Emerald Cloud Lab provides a software interface to remotely control every aspect of experiments run in ECL's automated facilities

1. COMMAND

Ship your samples to ECL for lab work. Then, design your experiments in the ECL Command Center application



2. EXECUTE

ECL conducts your experiments exactly to your specifications

3. EXPLORE

ECL stores deep, linked data from your experiments into your own personal cloud database

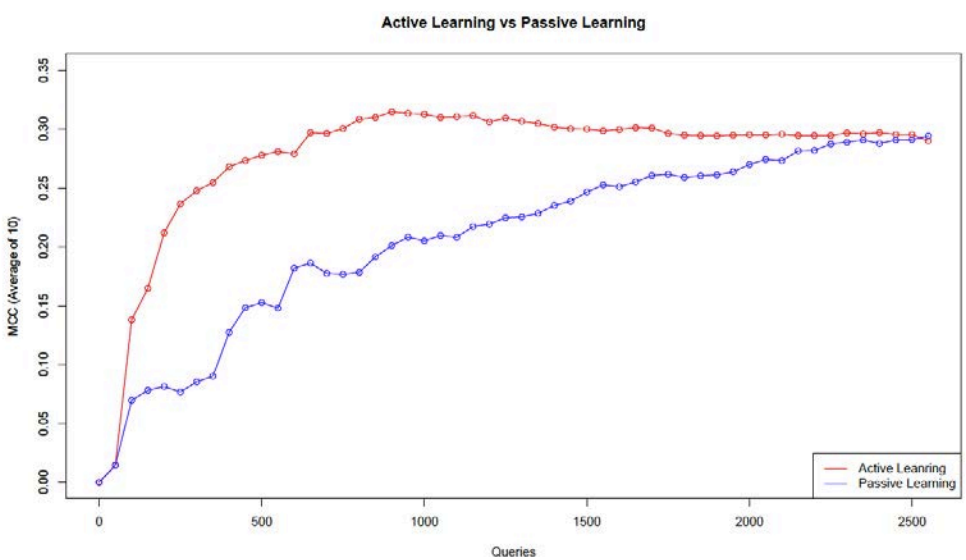
4. ANALYZE

Visualize & manipulate your results data with a powerful suite of the ECL Command Center tools

© 2017 Emerald Cloud Lab. All rights reserved.

Interpretable+Active learning (with oversight)

- Interpretable: explain why ML decisions are made
- Active: select new experiments that best improve the model (iterative)
- Oversight: Expert can veto "hypothesis" and "proposed experiment"—quantify value added of human in loop



Richard Philips '18 (→Cornell)

Example proposed experiment

```
* inorg1        : Oxovanadium(2+) sulfate
* inorg1mass    : 0.1648 grams
* inorg2        : Selenium dioxide
* inorg2mass    : 1.2995 grams
* org1          : 1,4-dimethylpiperazine
* org1mass      : 0.1166 grams
* temp          : 110
* time          : 24.0
* slowCool      : True
* pH            : 4
```

Uncertainty explanation

```
* 1.76 < PaulingElectronegGeom <= 1.89
* 1.76 < PaulingElectronegMean <= 1.91
* orgminimalprojectionradiusMax <= 2.84
* slowCool = True
* orghbdamsaccGeomAvg <= 0.99
* numberInorg = 2
```

Lessons learned

- "Failed" reactions make better classifiers…
  - Find overlooked "surprises" latent in archival data
  - *Challenge*: Need for public repositories/datasets
- …simple models can be useful…
  - Increased laboratory productivity/novelty
  - *Challenges*:
    - meaningful representations of materials & experiments
    - small and wide data, with lots of human bias
- …and you might even learn something!
  - Keep us honest about our "folk-theories"
  - Gives us ideas of new ideas to test…
  - …and ideas on how to test them.
- Robots! Coming to a lab near you…
  - Not just speed…*reproducibility*….
  - Importance of *modularity*

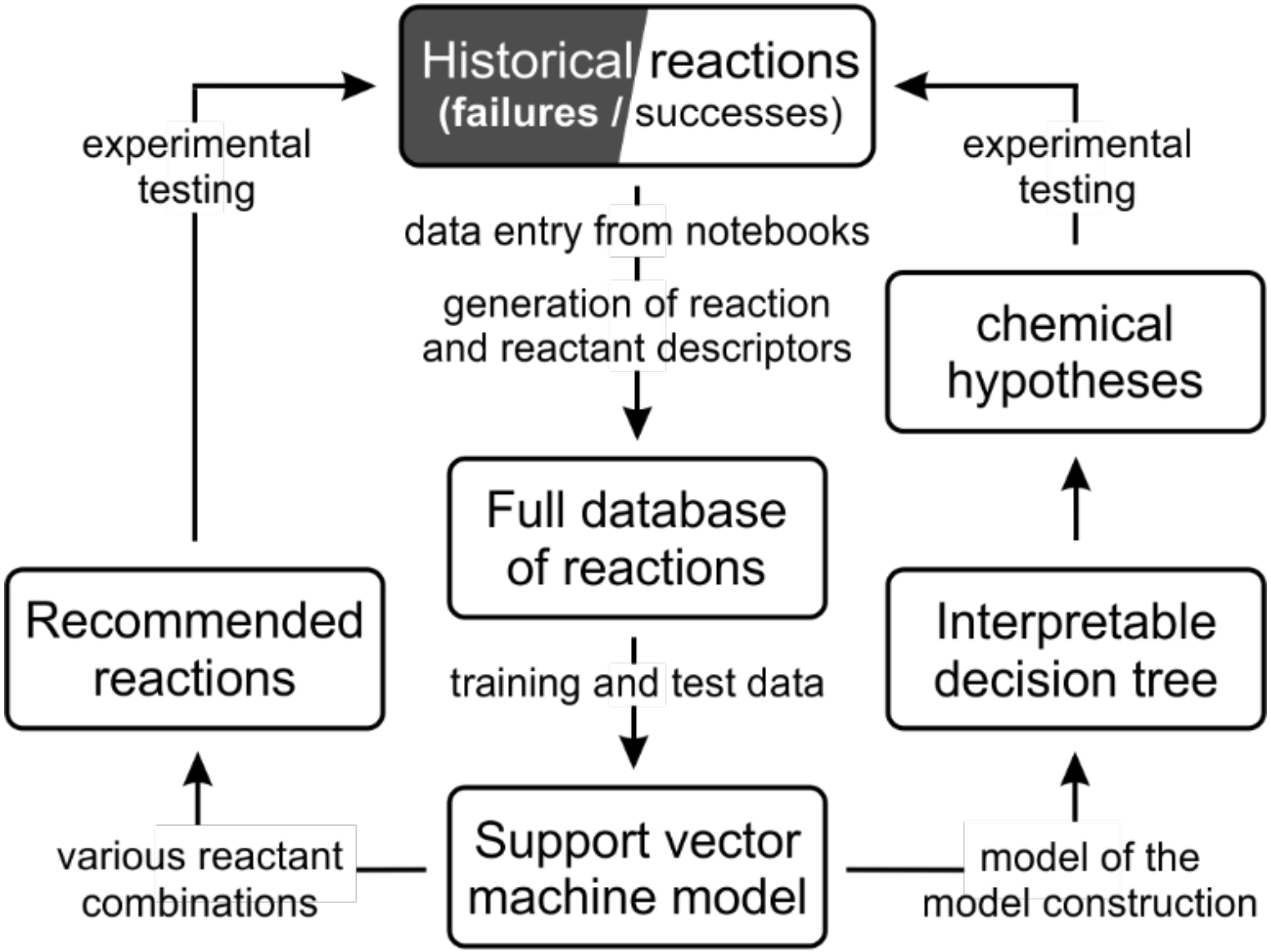
Acknowledgements

- Postdoctoral Fellows
  - Philip Adler (2015-2016) (Visa)—"original" dark reactions project
  - Ian Pendeton (2018-) — perovskites
- Undergraduate Students
  - Paul Raccuglia '14 (Google)—model development
  - Katie Elbert '14 (UPenn)—testing reactions
  - Yongjia "Scott" Huang '15 (JP Morgan/Chase) —testing reactions
  - Yunwen "Helen" Yang '15—testing reactions
  - Casey Falk '16 (Amazon) —website/database development
  - Geoffrey Martin-Noble '16 (Google)—Data entry, model development
  - David Reilley '16 (UCLA)—Data entry
  - Aurellio Mollo '17 (Harvard)—testing reactions, DFT for new descriptors
  - Malia Wenny '17 (Harvard)—testing reactions, DFT for new descriptors, initial perovskites work
  - Philip Nega '16 (UPenn)—testing reactions, experiment design project
  - Brian Guggenheimer '16 (Amazon)—website/visualizations
  - Nora Tien '18 (Google)—website/visualizations, model development
  - Rosalind Xu '18 (→Harvard)—testing reactions, factorial expt design project;
  - Xiwen Jia '19 —structural outcome descriptors; irrationality in experiment planning
  - Richard Philips '19 (→Cornell)—active learning
  - Liana Alves '18 (→ UC San Diego) —robotic synthesis of perovskites
  - Peter Cruz Parilla '20—hypothesis validation, perovskite chemistry
  - Alyssa Sherman '18 (→UPenn) — hypothesis validation, perovskite chemistry,

Acknowledgements

- Collaborators
  - Emory Chan (Molecular Foundry, LBNL)—high-throughput perovskite synthesis
  - Sorelle Friedler (Haverford)—interpretable models
  - Alexander Norquist (Haverford)—hydrothermal synthesis
  - Emerald Cloud Labs—robot army

- Funding
  - NSF DMR-1307801 & DMR-1709351
  - DARPA HR001118C0036
  - Work at Molecular Foundry: DOE BES DE-AC02-05CH11231
  - DOE Workforce Development (WDTS) Visiting Faculty Program (VFP)
  - Dreyfus-Teacher Scholar Award



Questions? jschrier@haverford.edu

http://darkreactions.haverford.edu