Learning patterns of chemical reactivity from experimental data

There's a lot to cover! Don't worry about writing down references; email me and I'll send you a copy of the slides!

Connor W. Coley
Assistant Professor
MIT Chemical Engineering
MIT Electrical Engineering and Computer Science

Cheminformatics Strasbourg Summer School
June 28, 2022

Klavs Jensen
William Green
Regina Barzilay
Tommi Jaakkola
Chris Voigt
Daniela Rus
Sangeeta Bhatia
Timothy Jamison

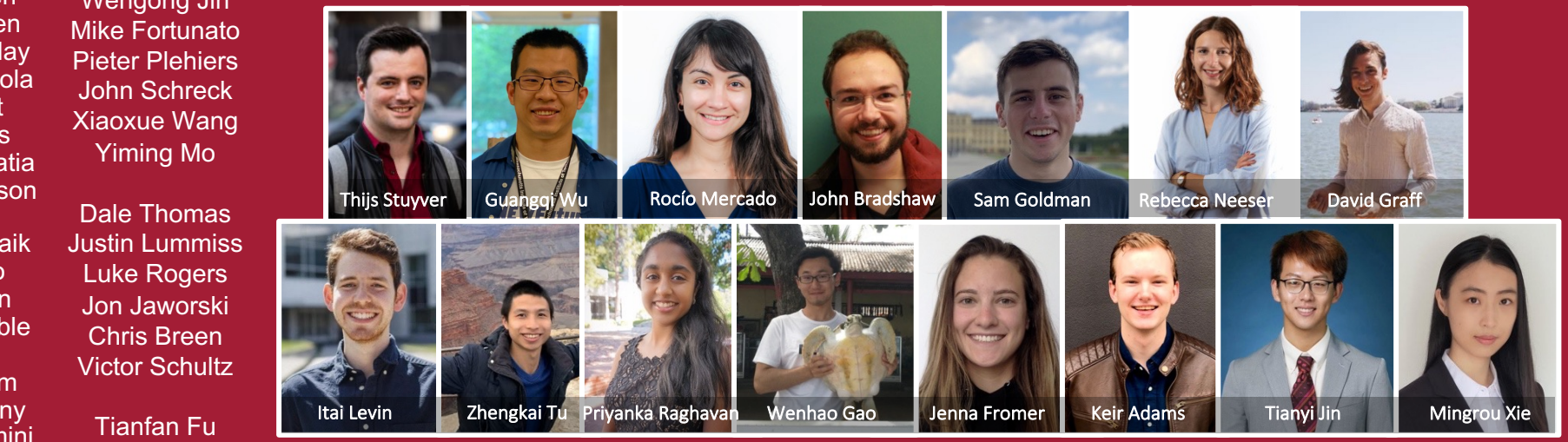
Lucky Pattanaik
Hanyu Gao
Yanfei Guan
Thomas Struble

Vignesh Ram
Ava Soleimany
Alexander Amini
Natalie Eyke

Wengong Jin
Mike Fortunato
Pieter Plehiers
John Schreck
Xiaoxue Wang
Yiming Mo

Dale Thomas
Justin Lummiss
Luke Rogers
Jon Jaworski
Chris Breen
Victor Schultz

Tianfan Fu
Jimeng Sun



+ Saul Vega Sauceda, Janet Li, Divya Nori, Chanwoo Yoon, Ron Shprints

Funding
MLPDS, DARPA, ONR, NIH, NSF, Takeda, Nanite Bio, J-Clinic, MIT-IBM Watson AI Lab

Data
CAS, Pistachio, Reaxys

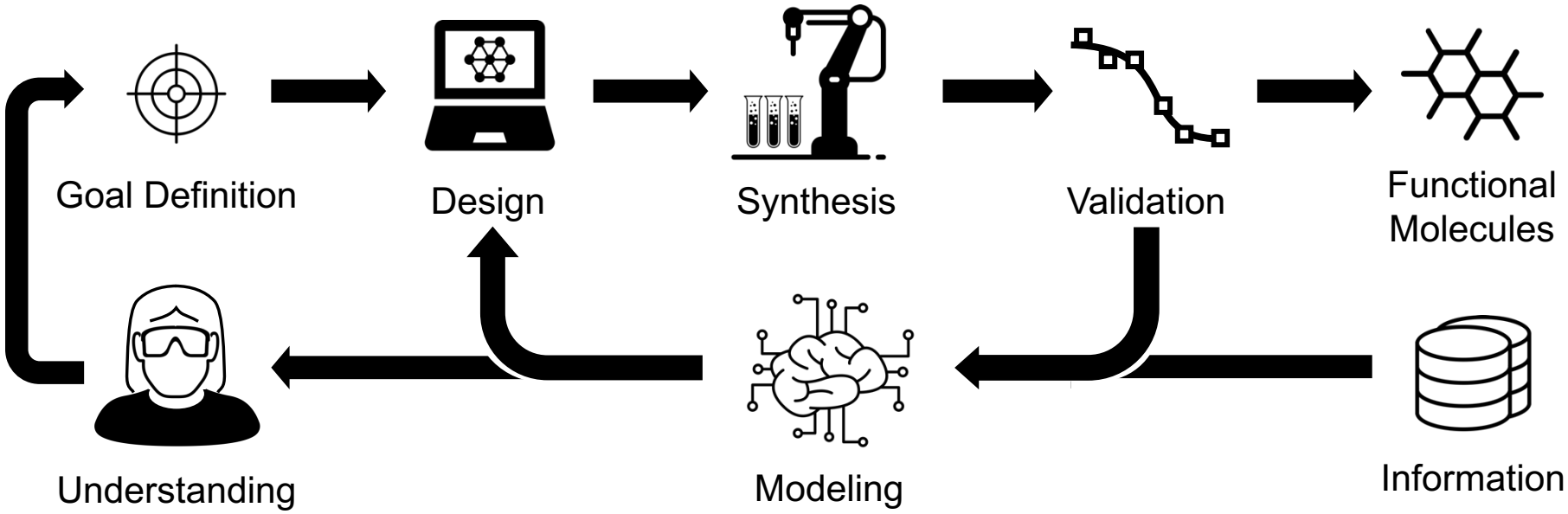
Advising
Entos, Galixir, Dow, Revela, Kebotix, Anagenex

ccoley@mit.edu | coley.mit.edu | @cwcoley

Closed-loop discovery & autonomous laboratories

e.g., small molecule hit-to-lead and lead optimization

Challenge: Molecular discovery is labor-intensive and heavily biased by human intuition. It takes immense resources and many years to bring a single drug to market; materials often take decades.



Overarching goal: Enable the "self-driving laboratory" through data-informed decision making and *information-centric* discovery

1. Domain-tailored neural models

2. Computer-aided molecular design

3. Data-driven predictive chemistry

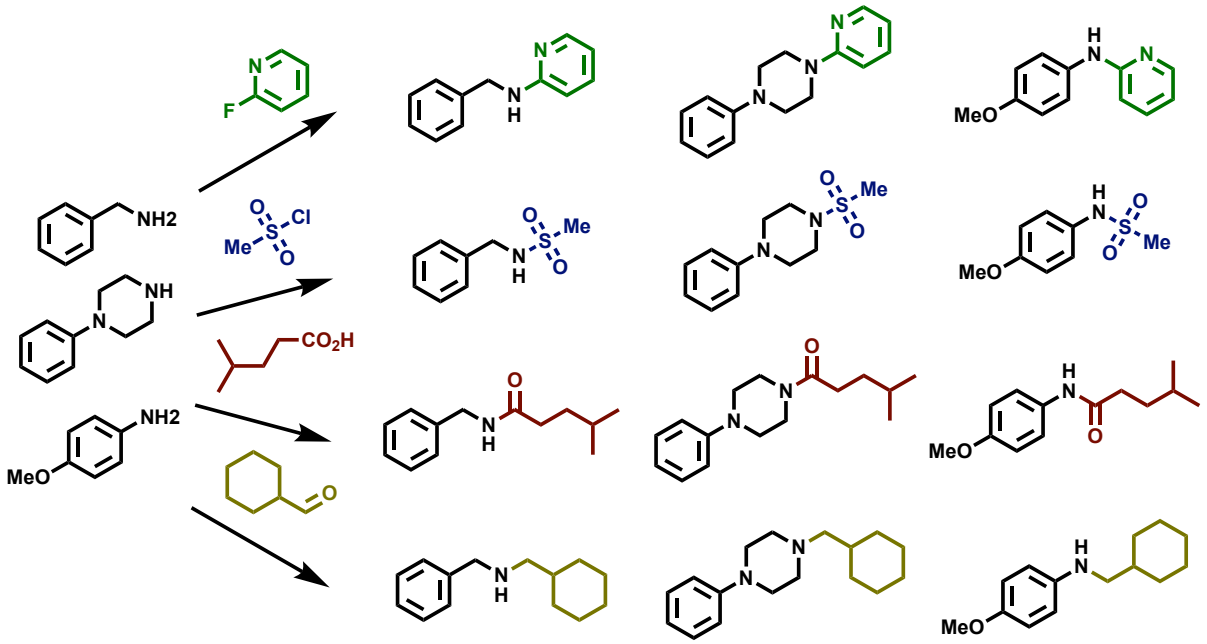
4. Robotics and laboratory automation

Relevance of reactivity to molecular design

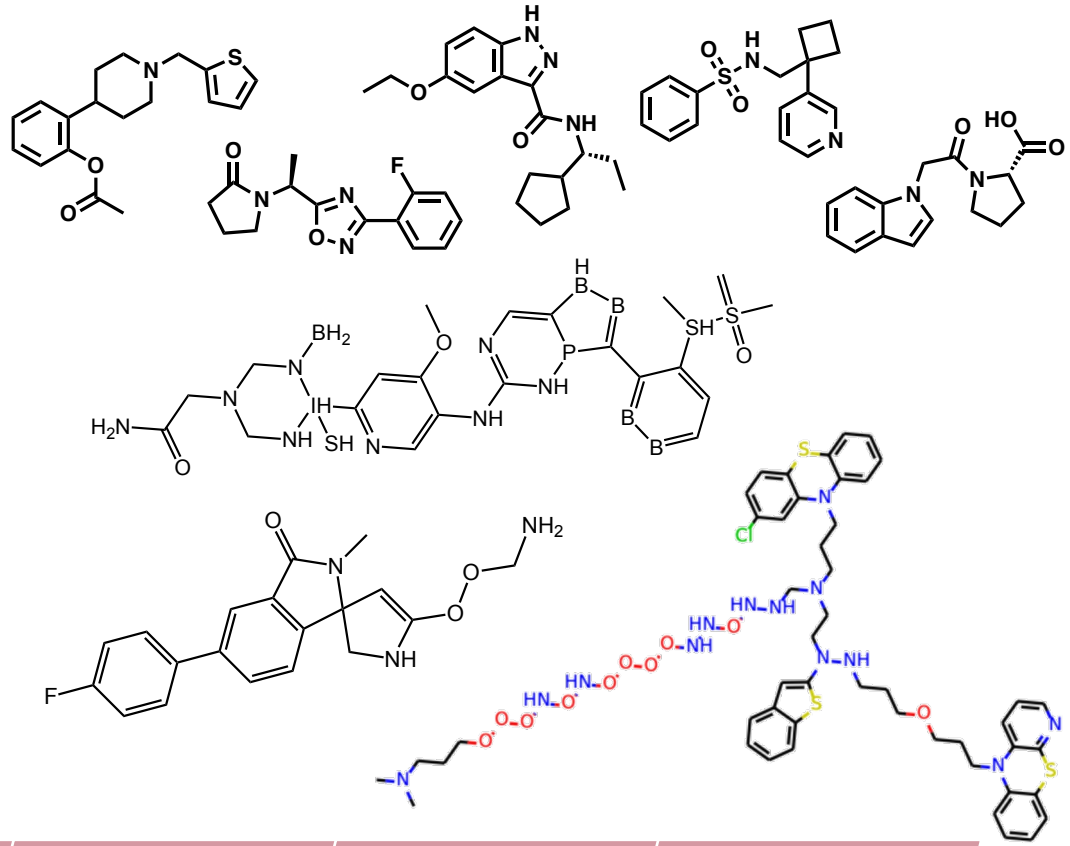
Synthesis constrains what we can access *at all* and influences what we can access *easily*

Coley, *Trends Chem.* **2021**.

Virtual libraries are often "make-on-demand" libraries enumerated using chemical transformation rules we believe to be robust



Generative models produce new compounds for which we must plan synthetic routes



Predictive chemistry (reaction informatics) tasks

Primary learning objective for this talk: understand the basics of reaction datasets, representation considerations (beyond molecular representation considerations), *and the common learning tasks*

Deployment

Retrosynthesis, reaction product prediction, classification/mapping

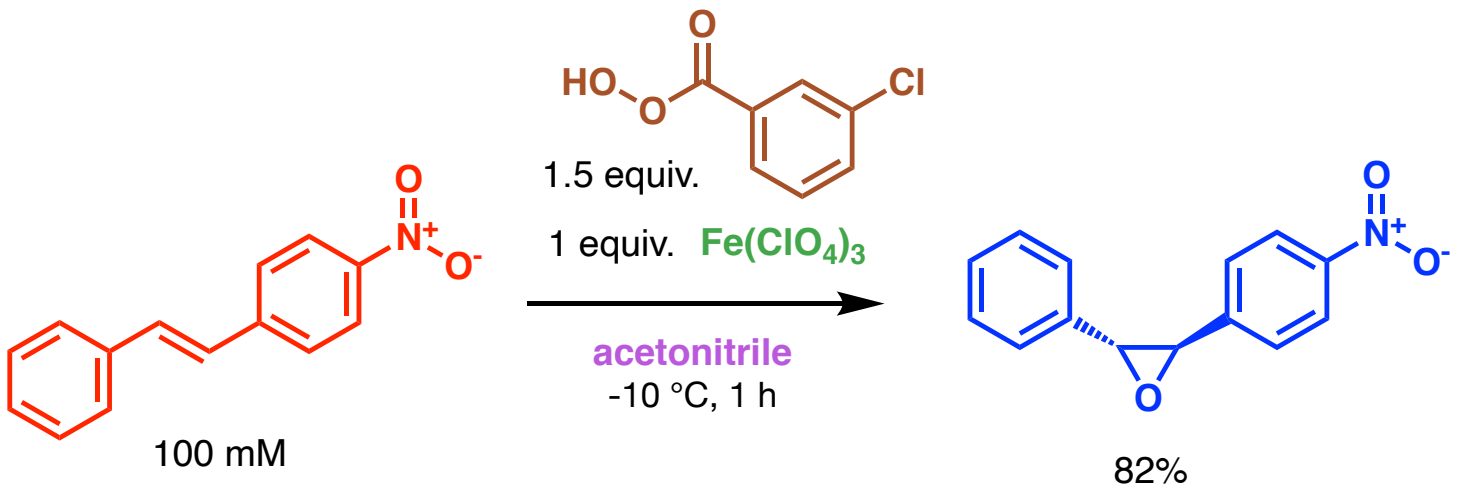
Development

Condition recommendation, condition optimization, scope assessment, catalyst design

Discovery

Mechanistic elucidation, new method development

Reactions as a data structure: important concepts



Quantitative aspects (concentrations, temperature; time; yield) and roles of agents are lost in line notations

RDFiles are “most general” as extension of SDFiles, but additional data does not have a universal format

Reaction SMILES:

```
O=[N+]([O-])C1=CC=C(C=C1)/C=C/C2=CC=CC=C2>>[O-][N+](C(C=C1)=CC=C1[C@H]2O[C@@H]2C3=CC=CC=C3)=O
```

Reaction SMILES with conditions:

```
O=[N+]([O-])C1=CC=C(C=C1)/C=C/C2=CC=CC=C2>ClC1=CC(C(OO)=O)=CC=C1.[O-]Cl(=O)(=O)=O.[O-]Cl(=O)(=O)=O.[O-]Cl(=O)(
=O)=O.[Fe+3].CC#N>[O-][N+](C(C=C1)=CC=C1[C@H]2O[C@@H]2C3=CC=CC=C3)=O
```

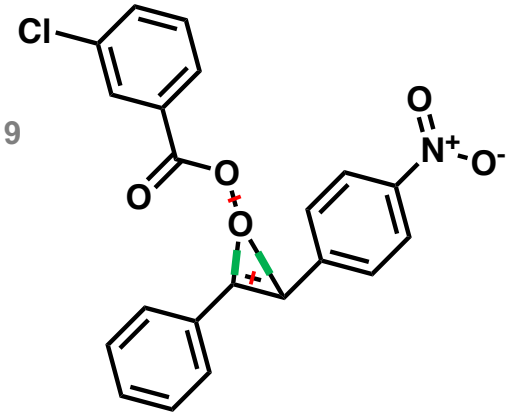
Atom-mapped reaction SMILES:

```
[O:3]=[N+:2]([O-:4])[C:1]1=[CH:5][CH:6]=[C:7]([CH:16]=[CH:17]1)/[CH:8]=[CH:9]/[C:10]2=[CH:11][CH:12]=[CH:13][C
H:14]=[CH:15]2>>[O-:3][N+:2]([C:1]([CH:5]=[CH:6]3)=[CH:17][CH:16]=[C:7]3[C@H:8]4O[C@@H:9]4[C:10]5=[CH:15][CH:1
4]=[CH:13][CH:12]=[CH:11]5)=[O:4]
```

Reactions as a data structure: representations

How are reactions represented as inputs for learning algorithms? *It can depend whether the product is meant to be part of the input*

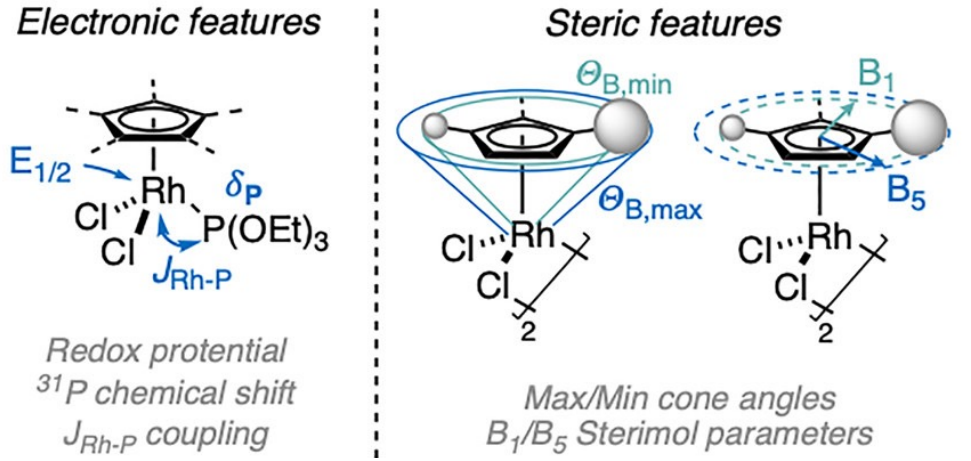
1. **Strings** (i.e., reaction SMILES) Schwaller et al. *Mach. Learn.: Sci. Technol.* 2, 015016, **2021**
2. **Constituent molecular components**
   a. Concatenation of reactant molecules (if fixed # components in dataset)
   b. Set of reactant molecules (if variable # components in dataset)
   c. Reaction difference fingerprints Schneider et al. *JCIM* 55(1) 39-53, **2015**
3. **Graphs & graph edits** (edits require atom-mapping) Coley et al. *Chem. Sci.* 10, 370-377, **2019**
4. **Condensed graph of reaction** (ignores spectators, requires atom-mapping)
   Hoonakker et al. *Int. J. Artif. Intell. Tools* 20(2) 253-270, **2011**; Heid and Green, *J. Chem. Inf. Model.* 62(9) 2101-2110, **2022**



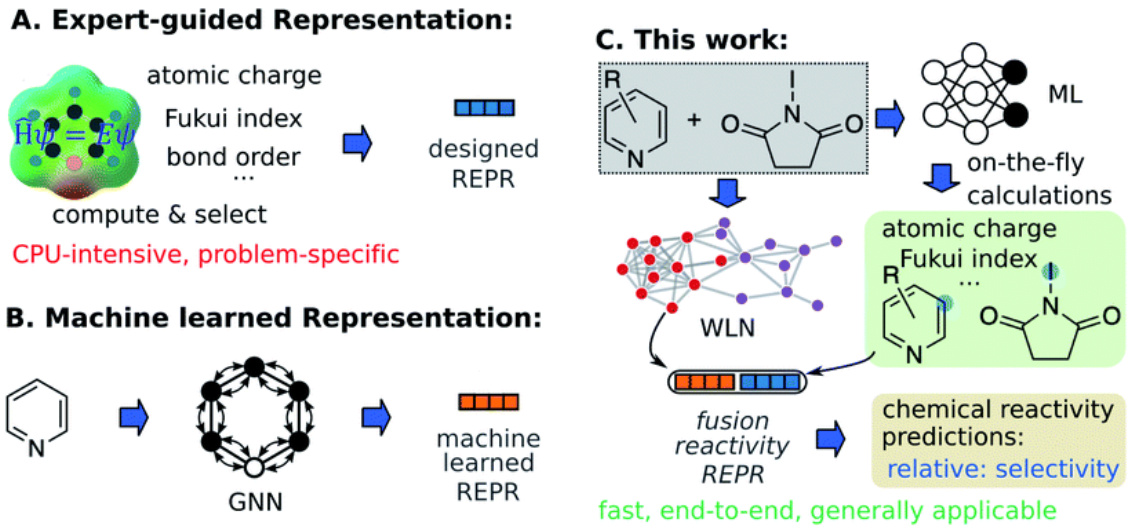
There is very little standardization in representing reaction conditions; often, the structures of catalysts, reagents, solvents are added to the reactants

Relationship to molecular representations

- Because reactions can always be represented by their constituent molecules, there is a very close relationship between *reaction representations* and *molecular representations*
- The **descriptor-based** v. **structure-based** "debate" applies to reactions as well
- Descriptors for complex catalysts are still very common, given the limitations of SMILES and graphs to describe complex catalysts or capture subtle aspects of structure



Gallegos et al. *Acc. Chem. Res.* 54(4) 827-836, **2021**



Guan et al. *Chem. Sci.* 12, 2198-2208, **2021**

Reaction data sources (select examples)

Patents	Literature	High-throughput Experimentation
USPTO (CC0 licensed)	Daniel Lowe, https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873	
Pistachio (commercial)	NextMove Software, https://www.nextmovesoftware.com/pistachio.html	
Reaxys (commercial)	Elsevier, https://www.reaxys.com/	
CAS / SciFinder (commercial)	CAS, https://www.cas.org/cas-data/cas-reactions	
	Ahneman et al. *Science* 360(6385) 186-190, **2018**	Merck's C-N coupling data
	Perera et al. *Science* 359(6374) 429-434, **2018**	Pfizer's Suzuki data
		…
Open Reaction Database	Kearnes et al. *JACS* 145(45) 18820-18826, **2021**	

Detailed look at a Reaxys entry

concentrations

reaction mechanisms

Published reaction data provides

byproducts

minor products

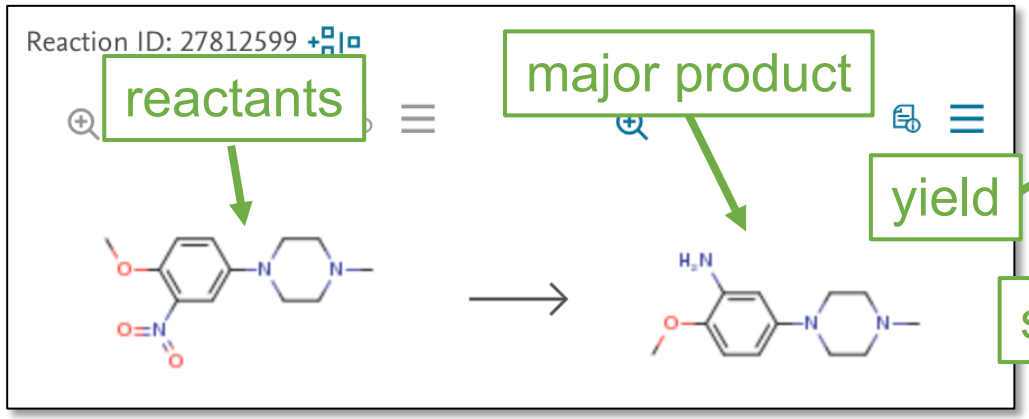
"failed" reactions

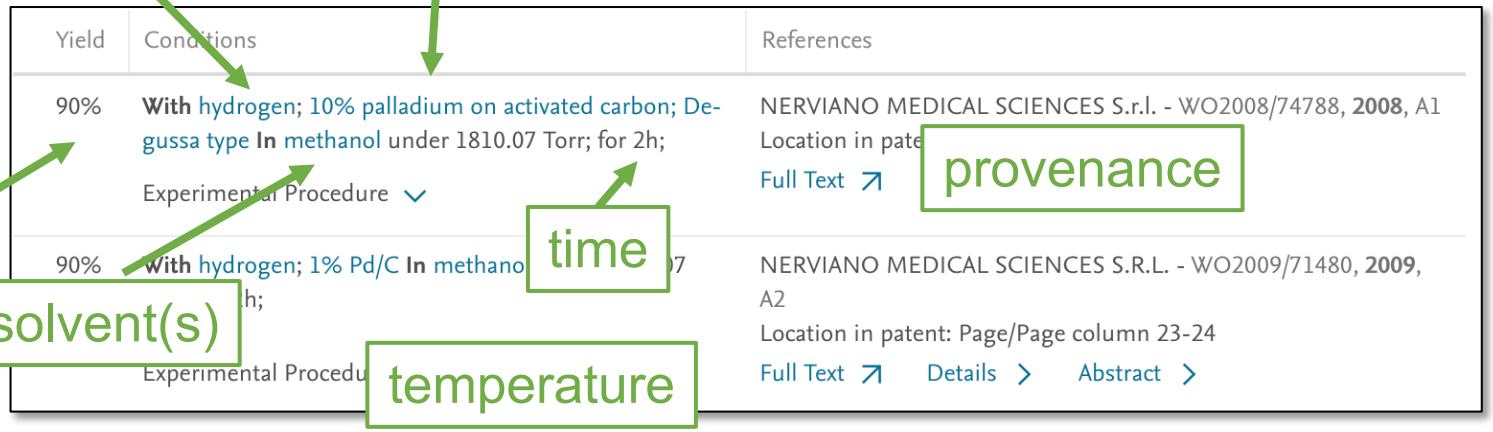
200+ years of experience to learn from

vessels

orders of addition

Millions of tabulated reaction examples





reagent(s)

catalyst(s)

yield

provenance

time

solvent(s)

temperature

Aside on evaluation & well-posedness

Models are typically evaluated in terms of their ability to recapitulate literature/experimental data

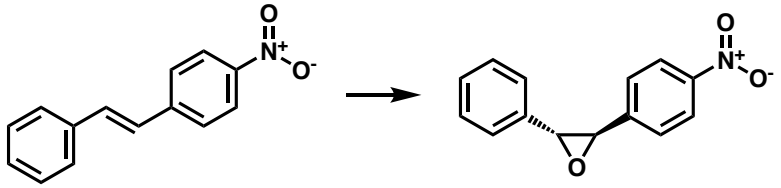
Any recommender system proposing new reactions cannot be evaluated with full confidence, since we cannot perfectly anticipate success/failure

- **One-step retrosynthetic analysis** always has more than one right answer
  - We do it anyway
- **Multi-step retrosynthetic analysis** has *many* more than one right answer
  - We (usually) evaluate models qualitatively, or in terms of their ability to find any pathway
- **Reaction outcome prediction** is underspecified without full knowledge of reaction conditions
  - We do it anyway
- **Yield prediction** is similarly underspecified when using literature data
  - We do it anyway

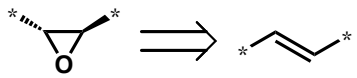
Computer-aided retrosynthetic analysis

Input: product molecule; **Output**: ≥1 reactant molecule(s)

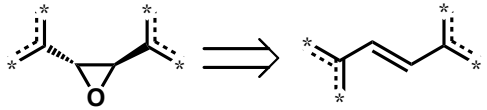
- Reaction templates can crudely codify the "rules of chemistry" Many prior publications; Coley et al., *J. Chem. Inf. Model.* 59, **2019**



1 **Find core of transformation**



2 **Add generalized neighbors**

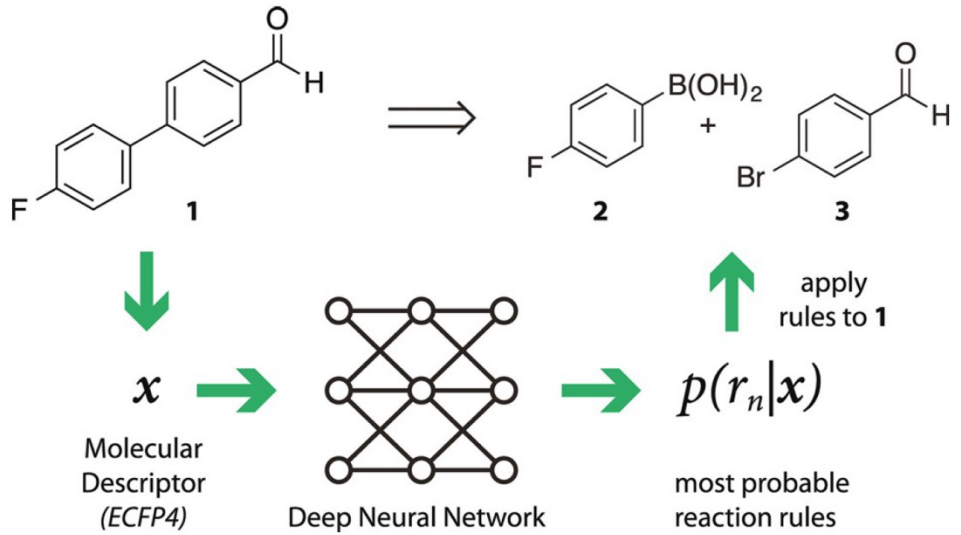


3 **Extend to known func. groups**

4 **Canonicalize and record**

```
[c:1]-[CH;@@;D3;+0:2]1-[O;H0;D2;+0]-
      [CH;@;D3;+0:3]-1-[c:4]
                >>
[c:1]/[CH;D2;+0:2]=[CH;D2;+0:3]/[c:4]
```

- After extracting a library of templates from a library of reactions, a classification model can learn when to apply them to new product molecules of interest Segler and Waller, *Chem. Eur. J. 23,* **2017**

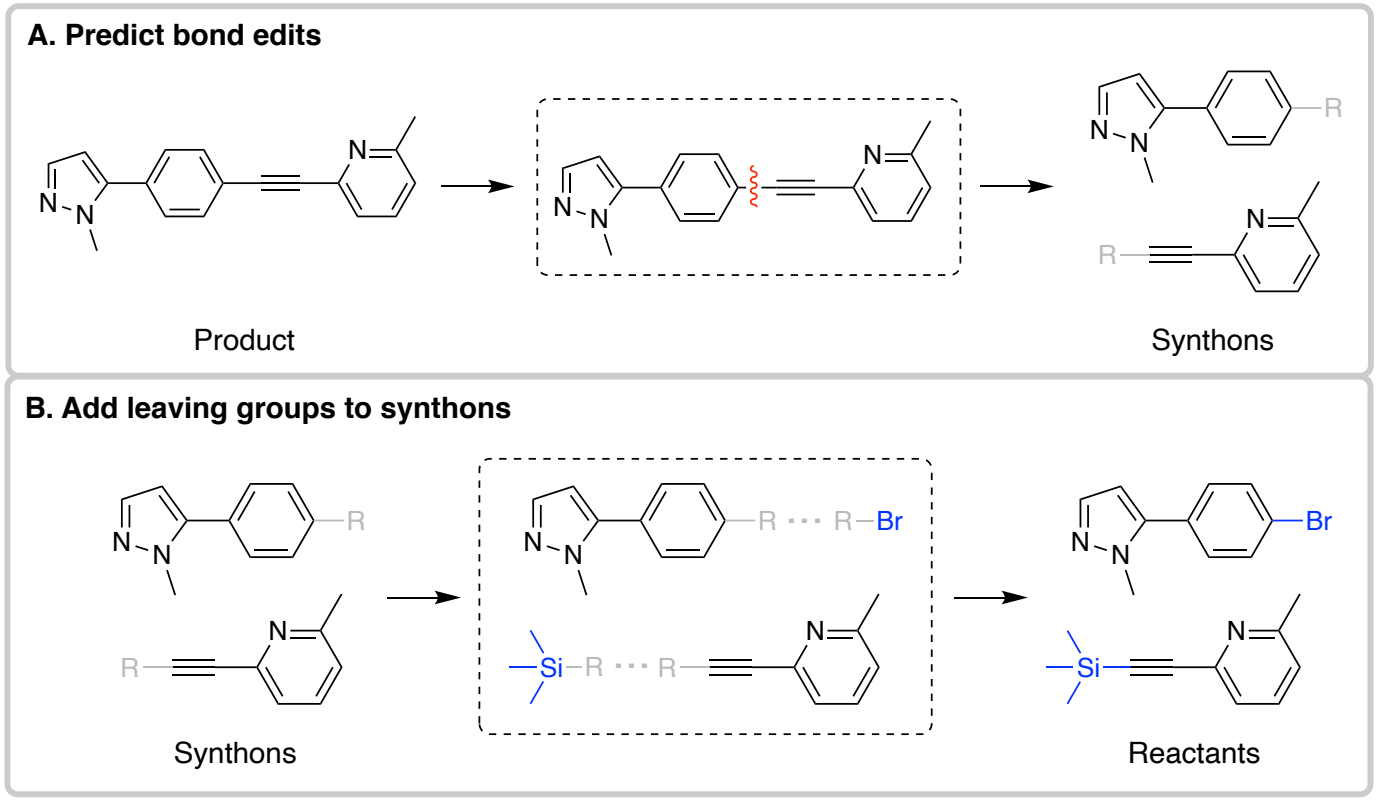


Low data approaches: Fortunato et al. *JCIM* 60(7) 3398-3407, **2020**; Seidl et al. *JCIM* 62(9) 2111-2120, **2022**

Computer-aided retrosynthetic analysis: template-free

Input: product molecule; **Output**: ≥1 reactant molecule(s)

- There are many template-free formulations of the one-step retrosynthesis task, including
  - **SMILES-to-SMILES** Liu et al., *ACS Cent. Sci.* 3, **2017**; Schwaller et al. *Chem. Sci.* 11, 3316-3325, **2020**; Lin et al., *Chem. Sci.* 12, **2020**; etc.
  - **Graph-to-Graph** Shi et al., *ICML* **2020;** Ram et al. *arxiv:2006.07038* **2020;** Sacha et al. *arxiv:2006.15426* **2020**



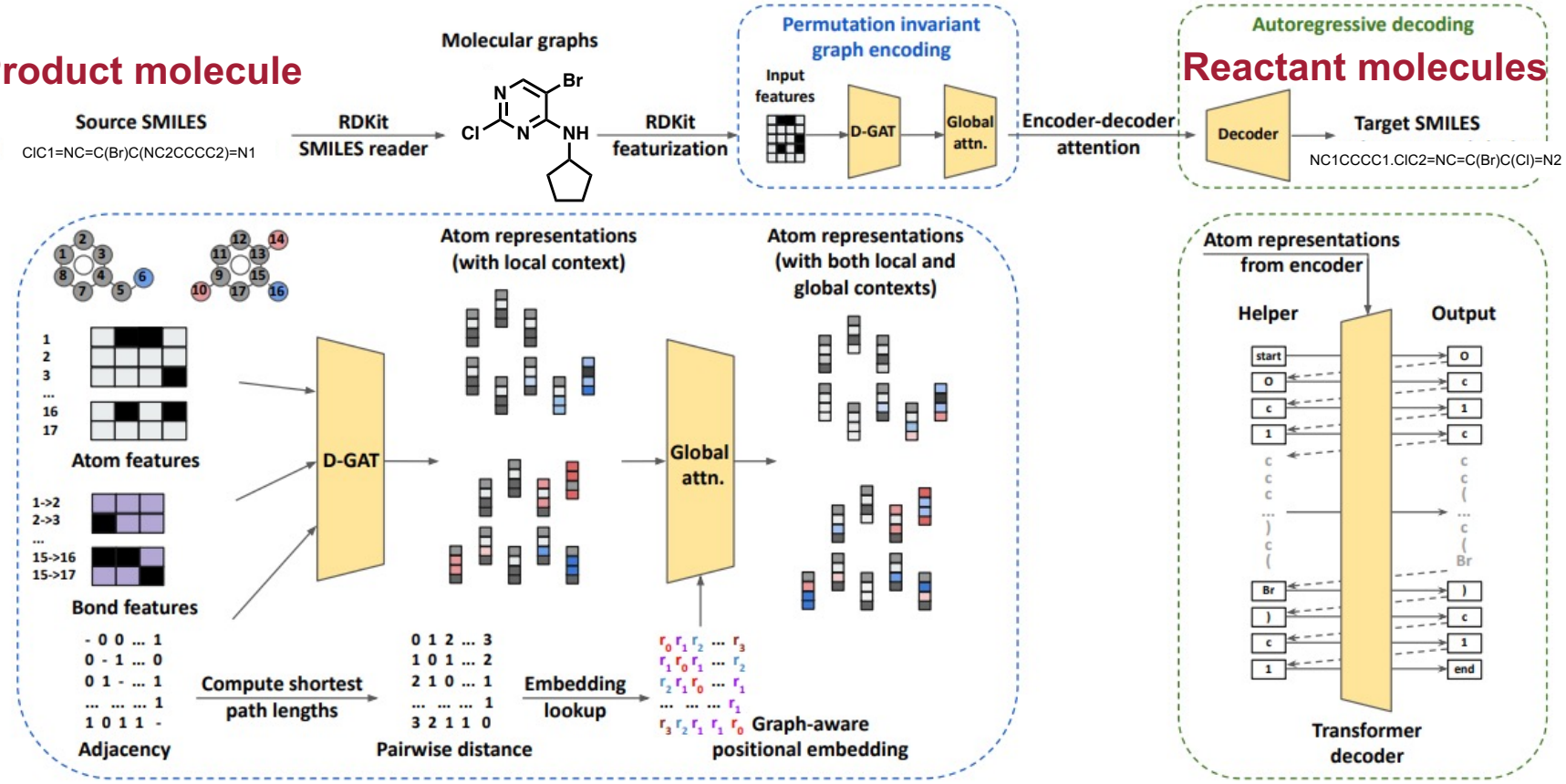
= bond selection

= graph completion
(or categorical prediction
of the leaving group)

Computer-aided retrosynthetic analysis: template-free

Input: product molecule; **Output**: ≥1 reactant molecule(s)

- There are many template-free formulations of the one-step retrosynthesis task, including
  - **SMILES-to-SMILES** Liu et al., *ACS Cent. Sci.* 3, **2017**; Schwaller et al. *Chem. Sci.* 11, 3316-3325, **2020**; Lin et al., *Chem. Sci.* 12, **2020**; etc.
  - **Graph-to-Graph** Shi et al., *ICML* **2020;** Ram et al. *arxiv:2006.07038* **2020;** Sacha et al. *arxiv:2006.15426* **2020**
  - **Graph-to-SMILES**



Tu and Coley, https://arxiv.org/abs/2110.09681

Computer-aided retrosynthetic analysis & evaluation

Input: product molecule; **Output**: ≥1 reactant molecule(s)

- Evaluation strategies focus on top-*n* accuracy, i.e., recapitulating literature examples
  - Most evaluations have focused on the "USPTO_50k" dataset, which is small and covers 10 classes
  - Some have looked at ca. 1M reactions from "USPTO_full"

Methods	Top-*n* accuracy (%)		Features / techniques used		
	1	10	*Templ.*	*Map.*	*Aug.*
RetroSim (Coley et al., 2017)	32.8	56.1	✓	✓	✗
MEGAN (Sacha et al., 2021)	33.6	63.9	✗	✓	✗
NeuralSym (Segler & Waller, 2017)	35.8	60.8	✓	✓	✗
GLN (Dai et al., 2019a)	39.3	63.7	✓	✓	✗
RetroPrime (Wang et al., 2021b)	44.1	68.5	✗	✓	✓
Aug. Transformer (Tetko et al., 2020)	44.4	**73.3**	✗	✗	✓
Graph2SMILES (D-GAT) (*ours*)	45.7	62.9	✗	✗	✗
Graph2SMILES (D-GCN) (*ours*)	45.7	63.4	✗	✗	✗
GTA (Seo et al., 2021)	**46.6**	70.4	✗	✓	✓

Tu and Coley, https://arxiv.org/abs/2110.09681

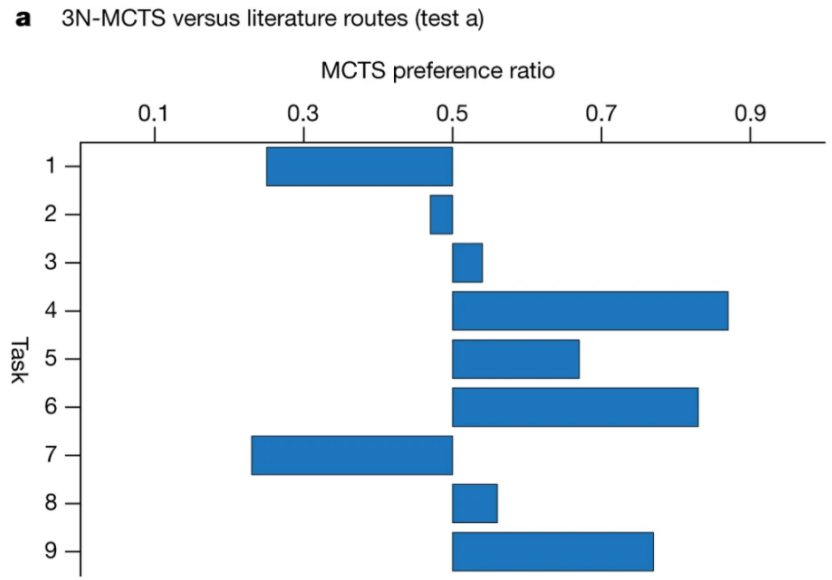
- Other evaluation metrics introduce additional assumptions, e.g., roundtrip accuracy

Schwaller et al., *ML4PhysicalScience @ NeurIPS* **2019** https://ml4physicalsciences.github.io/2019/files/NeurIPS_ML4PS_2019_116.pdf

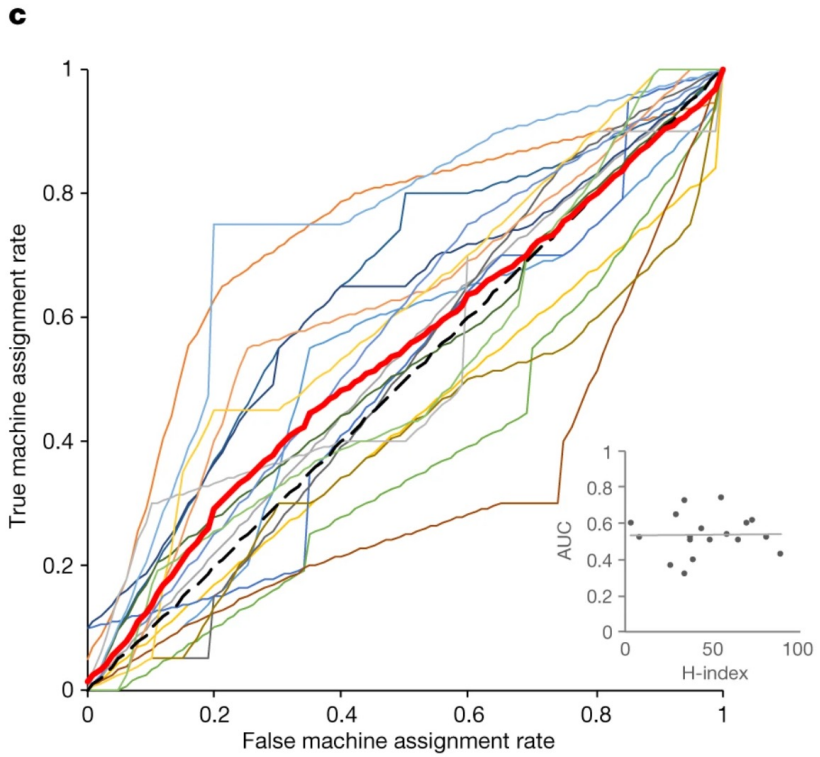
Multi-step planning

Input: product molecule; **Output**: synthetic pathway that terminates in buyable molecules

- Tree search strategies can perform recursive one-step expansion to connect complex molecules back to purchasable compounds (e.g., Monte Carlo tree search, best-first, proof number search)
- Both data-driven and expert approaches have generated pathways that score well in blinded tests:



Segler, Preuss, Waller
Nature 555, 604-610 (**2018**)

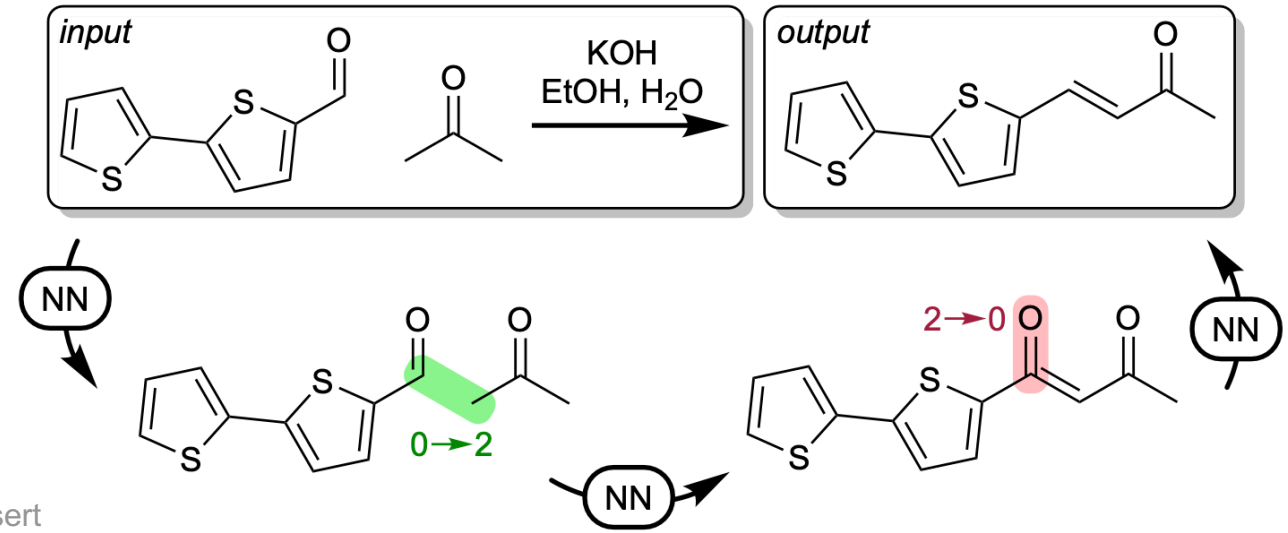


Mikulak-Klucznik, Grzybowksi et al.
Nature 588, 83-88 (**2020**)

Major product prediction

Input: ≥1 reactant molecule(s) + ≥0 agent molecule(s); **Output**: product molecule

- Mostly the reverse of retrosynthesis, except we don't have to create leaving groups from scratch
  - We can use graph edit methods that just rearrange bonds
- Template-free methods do not offer good coverage and generalizability Coley et al. *ACS Cent. Sci* 3(5) **2017**; Segler and Waller, *Chem. Eur. J.* 23, **2017**
- There are *also* many template-free formulations of the forward prediction task, including
  - **SMILES-to-SMILES** Schwaller et al., *Chem. Sci.* 9, **2018**; Schwaller et al., *ACS Cent. Sci.* 5, **2019**
  - **Graph-to-Graph** Jin et al. *NeurIPS* **2017**; Coley et al., *Chem. Sci.* 10, **2019**; Sacha et al. *JCIM* 61(7), **2021**; Bradshaw et al. *ICLR* 2019; …
  - **Graph-to-SMILES** Tu and Coley, https://arxiv.org/abs/2110.09681

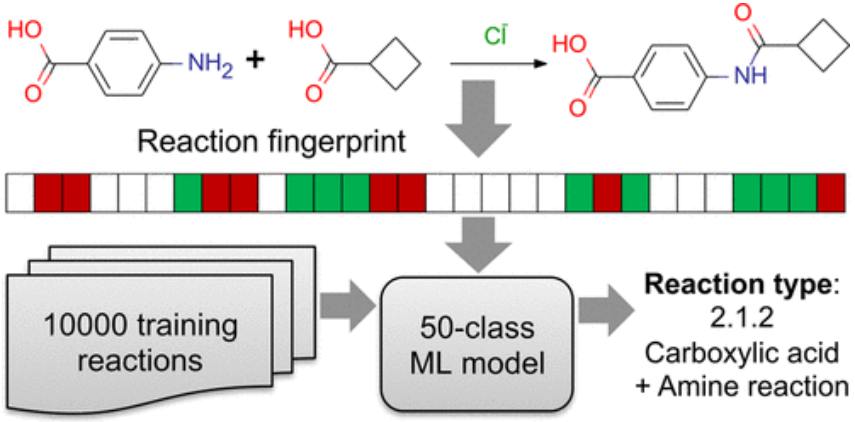


Clemens Isert

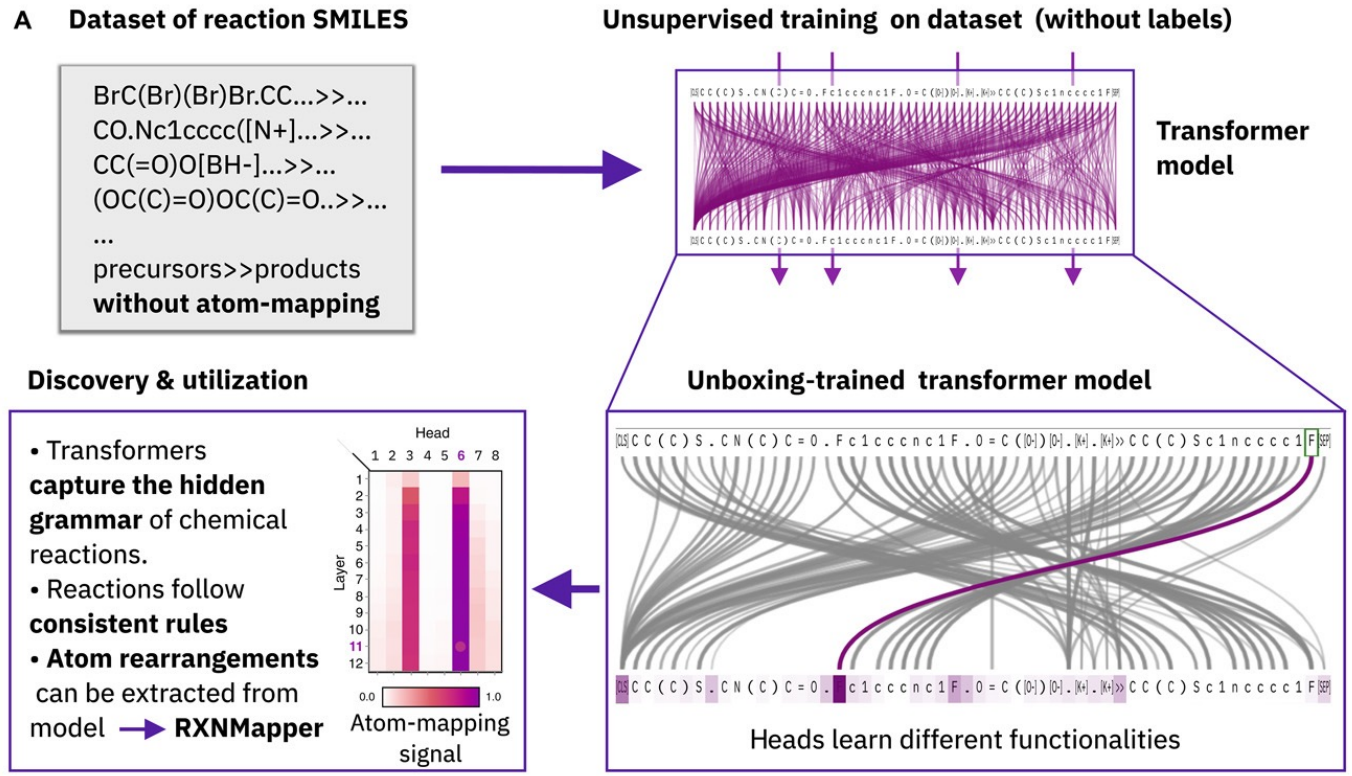
Classification & atom mapping

Input: ≥1 reactant molecule(s), ≥1 product molecule(s); **Output**: reaction type classification or atom-to-atom map

- Classically, these are rule-based or heuristic
  - E.g., maximum common substructures for performing atom mapping
  - E.g., NextMove Software's NameRXN is a set of SMARTS patterns that define reaction types
- Both tasks can be trained through supervised learning (reaction fingerprint, or even unsupervised learning from language models)



Schneider et al. *JCIM* 55(1) 39-53, **2015**

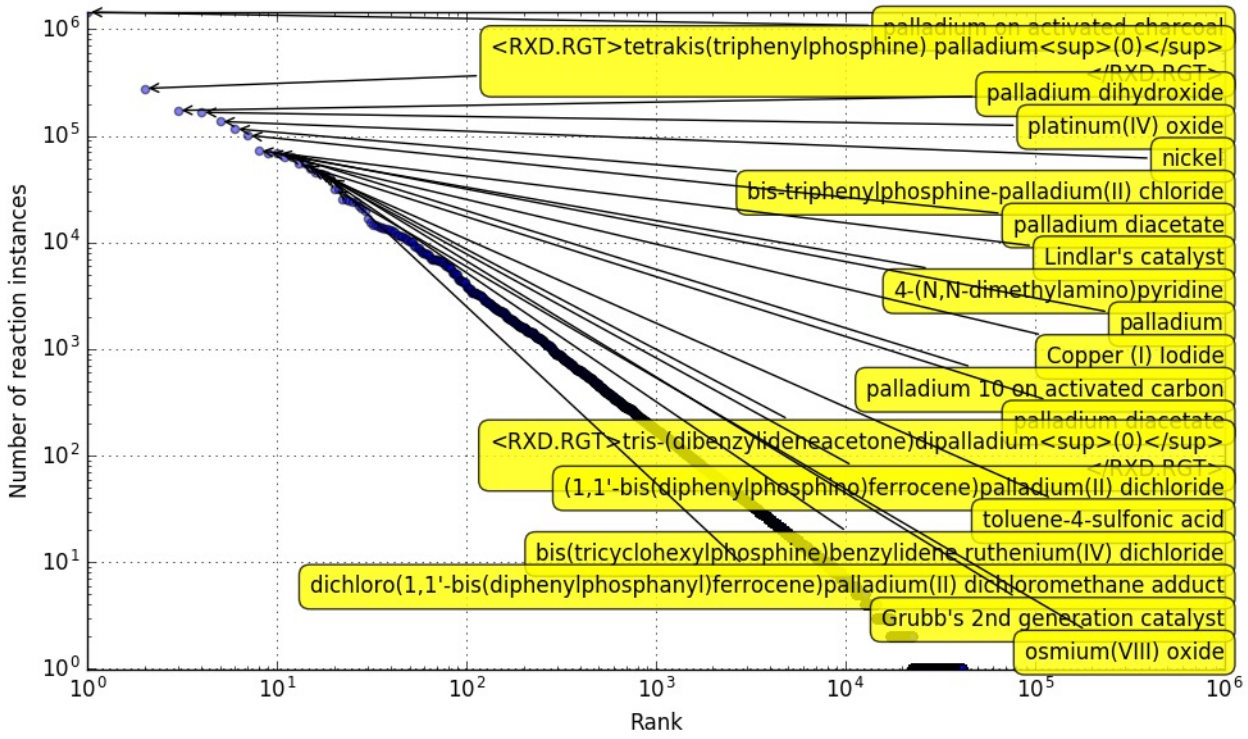


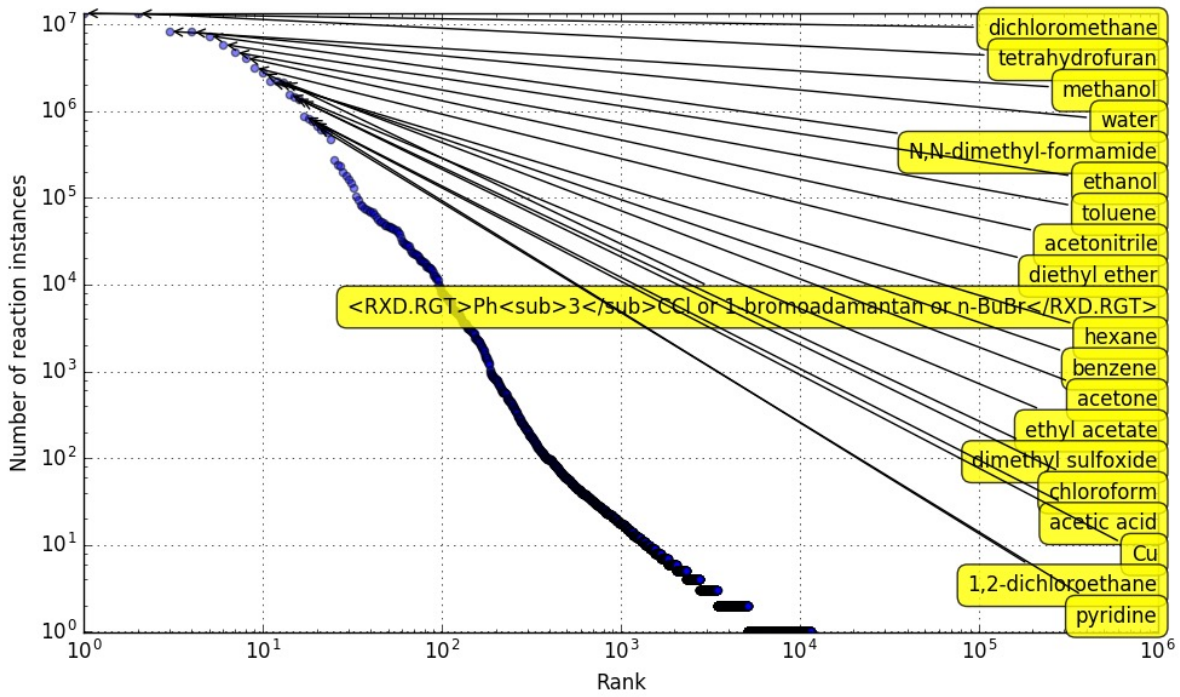
Schwaller et al. *Sci. Adv.* 7(15) **2021**

Reaction condition recommendation

Input: ≥1 reactant molecule(s), ≥1 product molecule(s); **Output**: varies

- Reaction conditions are essential for reaction execution; models to predict suitable conditions *a priori* can be trained on published data to “fill in the blank” above the arrow
- The modest number (thousands) of distinct reagents, catalysts, and solvents means that we can get away with a classification formulation rather than generation

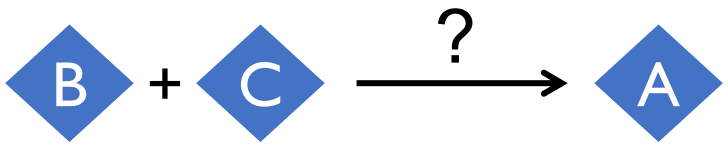




Reaction condition recommendation

Input: ≥1 reactant molecule(s), ≥1 product molecule(s); **Output**: varies

- Reaction conditions are essential for reaction execution; models to predict suitable conditions *a priori* can be trained on published data to "fill in the blank" above the arrow
- The modest number (thousands) of distinct reagents, catalysts, and solvents means that we can get away with a classification formulation rather than generation



- Multiple formulations have been pursued for data-driven condition prediction
  - Predict 1 catalyst, 1-2 solvents, 1-2 reagents, and temperature for "any" organic reaction as input
    Gao et al. *ACS Cent. Sci.* 4(11) 1465-1476, **2018**
  - Predict compounds for specific aspects of reaction conditions, for one reaction type at a time (e.g., metal, ligand, base, solvent, additive for Suzuki Maser et al. *JCIM* 61(1) 156-166, **2021**
  - Predict reagent-dependent yields and perform an *in silico* screen Nielsen et al. *JACS* 140(15) 5004-5008, **2018**

Model-guided reaction condition optimization

Input: ≥1 reactant molecule(s), ≥1 product molecule(s); detailed quantitative conditions **Output**: yield/performance

- If we have access to new experimental results for feedback, we can use surrogate model-guided optimization to propose improved concentrations, temperature, catalysts, etc.
- This is an old problem

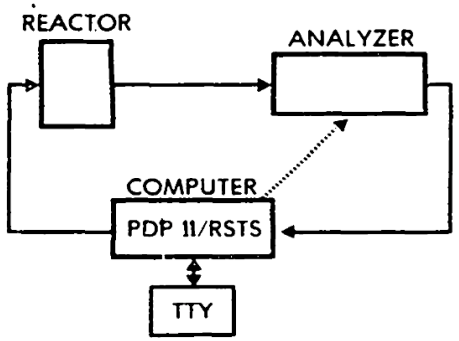


Fig. 1. Closed-loop system for automated chemical synthesis.

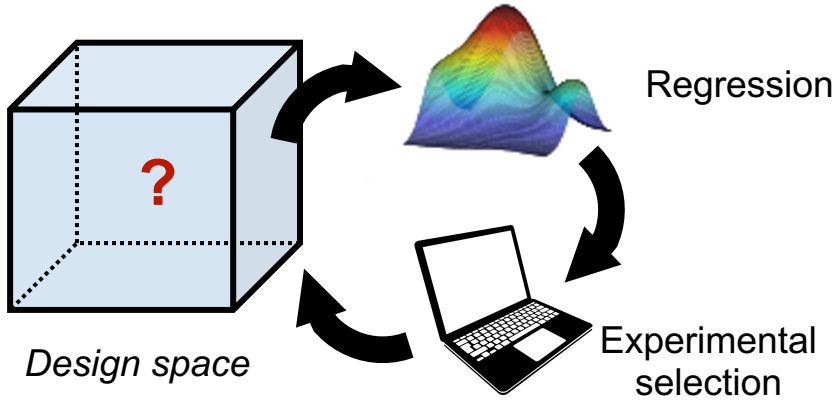
CHEMICAL PROCESS OPTIMIZATION BY COMPUTER — A SELF-DIRECTED CHEMICAL SYNTHESIS SYSTEM

H. WINICOV,* J. SCHAINBAUM, J. BUCKLEY, G. LONGINO, J. HILL and C. E. BERKOFF

Research and Development Division, Smith Kline and French Laboratories, Philadelphia, Pa. (U.S.A.)

(Received 3rd May 1978)

- More recent work by Ley, Jensen, Lapkin, Doyle, Bourne, etc. follows the same model-guided optimization workflow using a surrogate model to predict performance (e.g., yield)

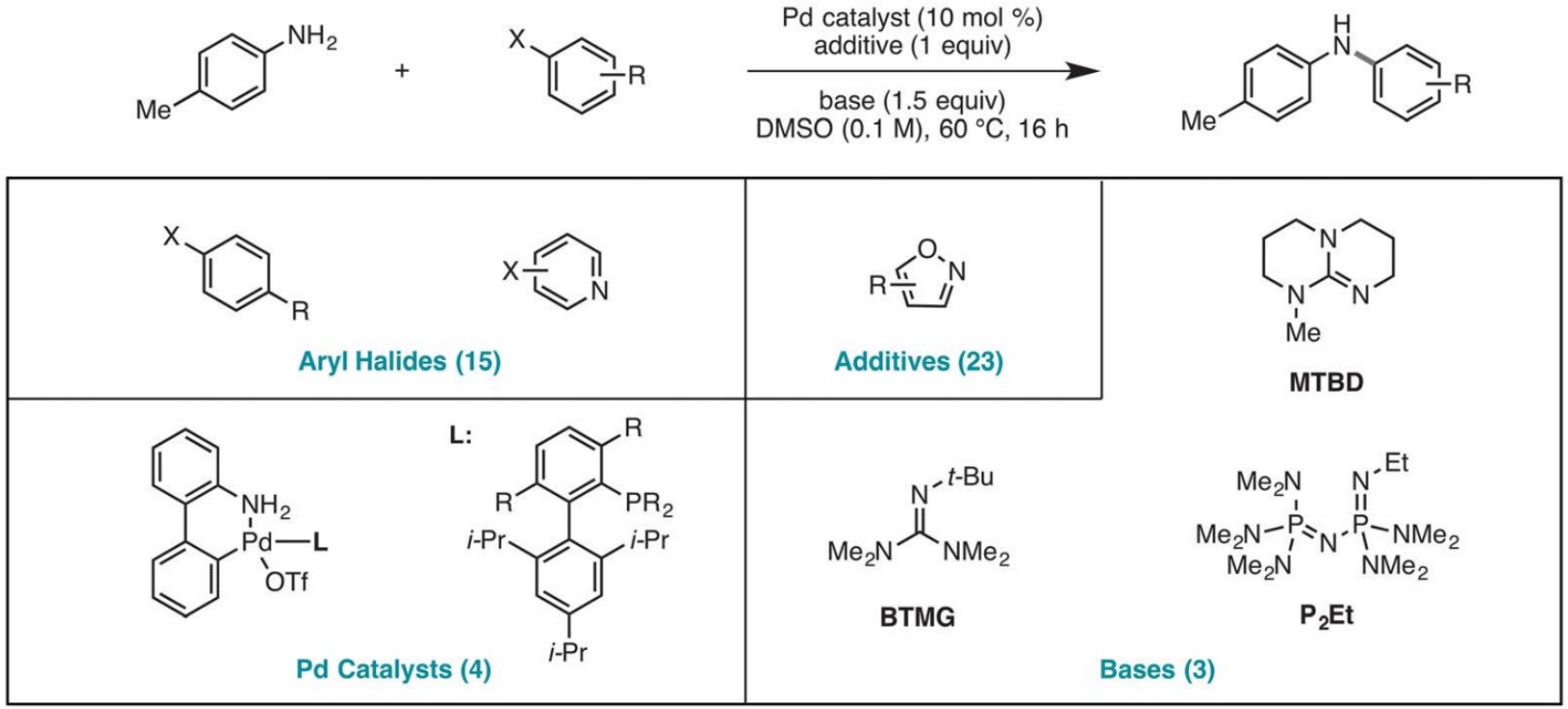


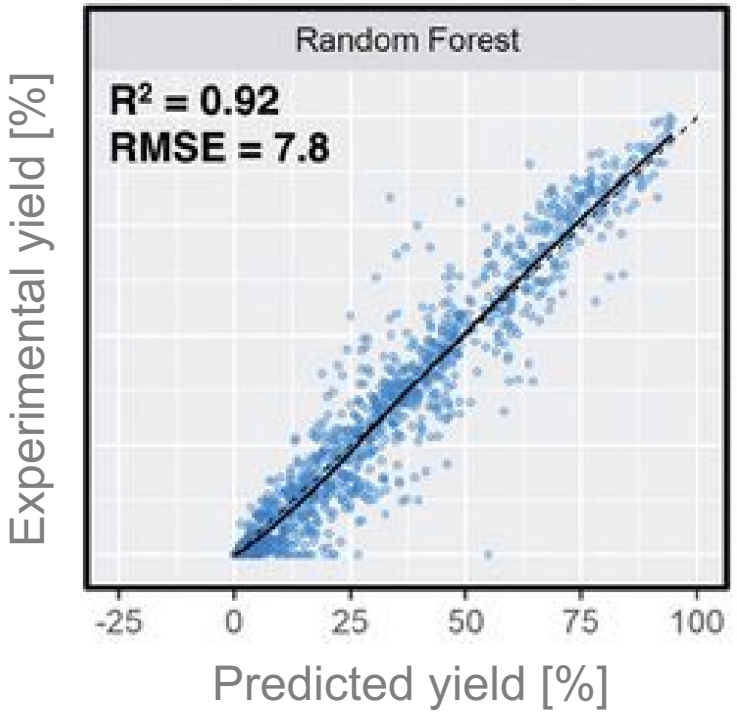
Substrate scope assessment & yield prediction

Input: hypothetical substrate; **Output**: yield or Boolean

- These models try to answer the question: “What substrates will work with my reaction?” Or, equivalently, “What substrates will lead to a good yield?”
- HTE provides a rich source of information where most aspects of the reaction are held constant

Ahneman et al. *Science* 360(6385) 186-190, **2018**



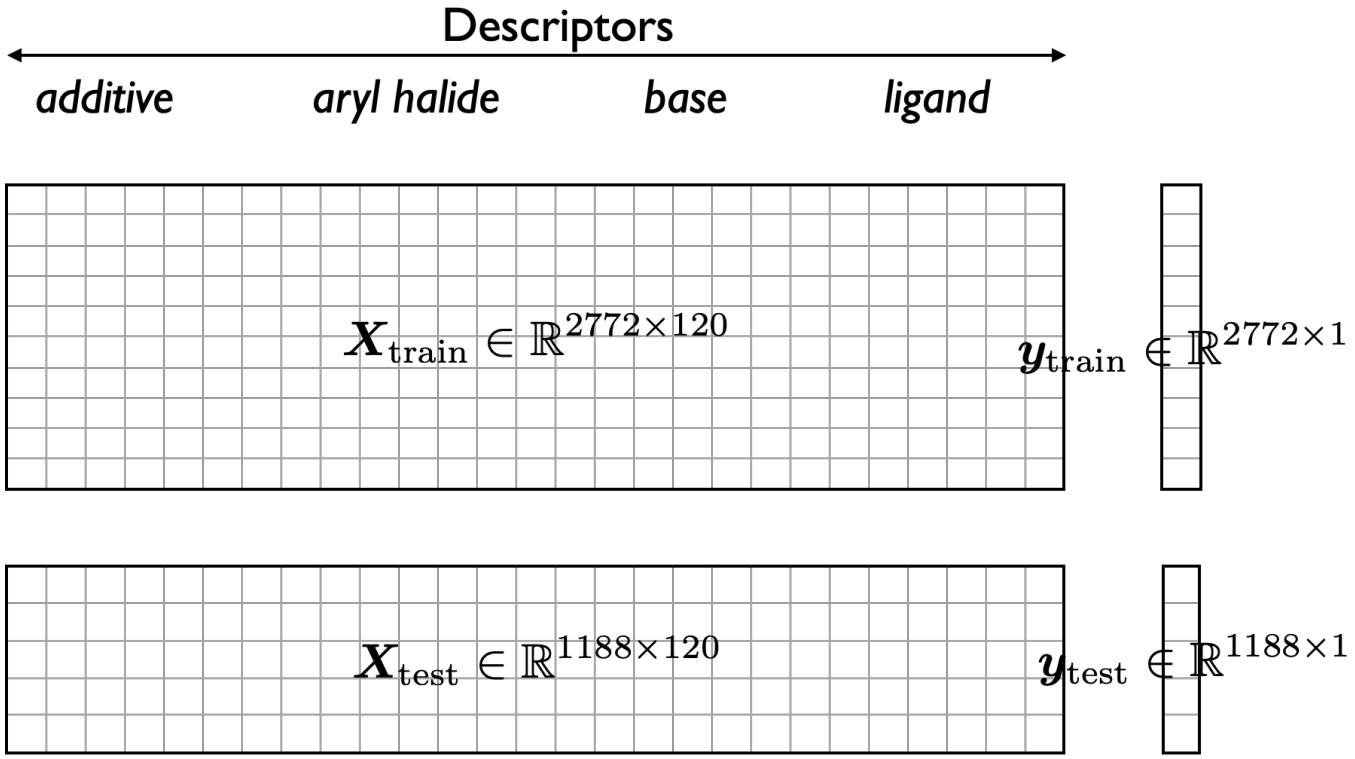


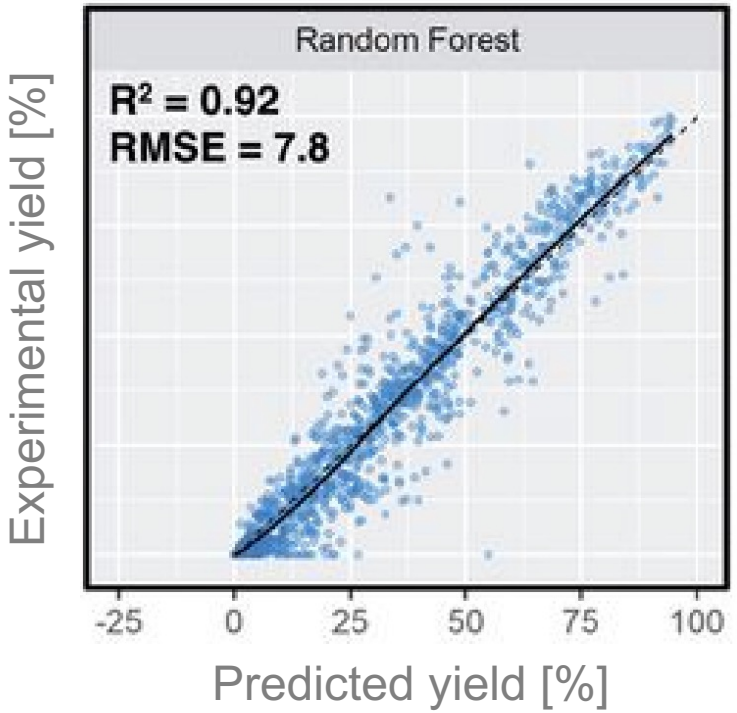
Substrate scope assessment & yield prediction

Input: hypothetical substrate; **Output**: yield or Boolean

- These models try to answer the question: “What substrates will work with my reaction?” Or, equivalently, “What substrates will lead to a good yield?”
- HTE provides a rich source of information where most aspects of the reaction are held constant

Ahneman et al. *Science* 360(6385) 186-190, **2018**





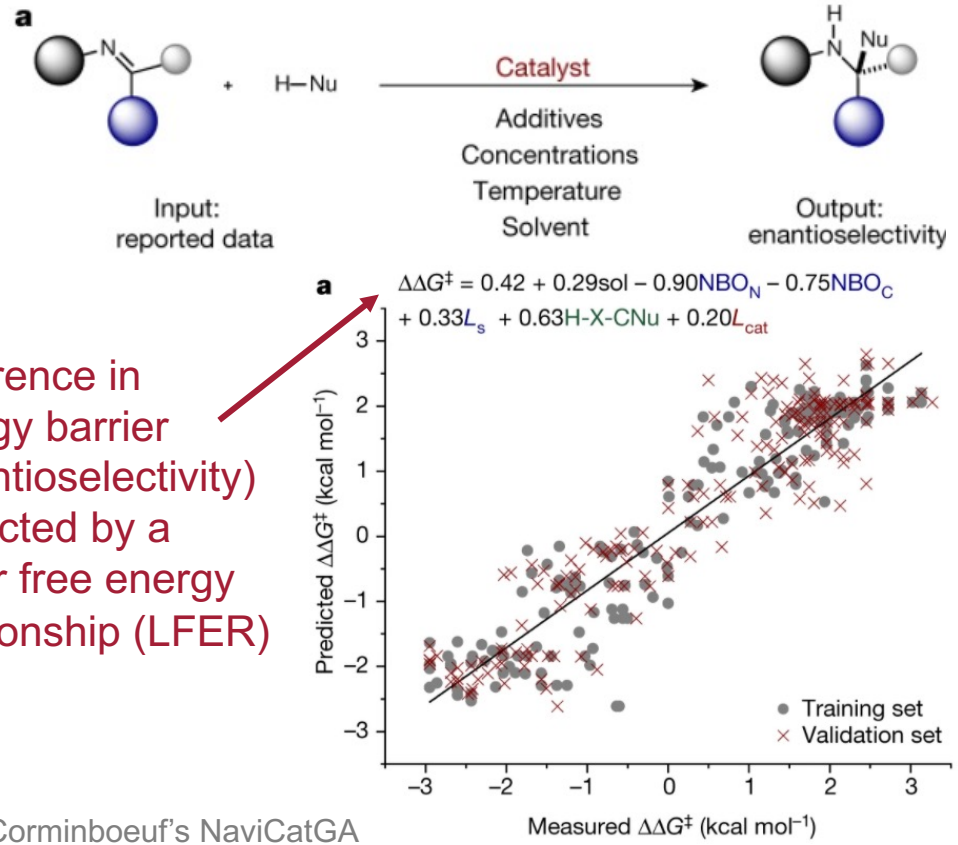
Model-guided catalyst/ligand design

Input: hypothetical catalyst/ligand structure or descriptors; **Output**: yield/selectivity/etc.

- Model-guided molecular design relies on models that look just like QSAR/QSPR: need to correlate structure to function, then we can virtually screen new (hypothetical) structures

Descriptor-based example

Reid & Sigman *Nature* 571, 343-348, **2019**

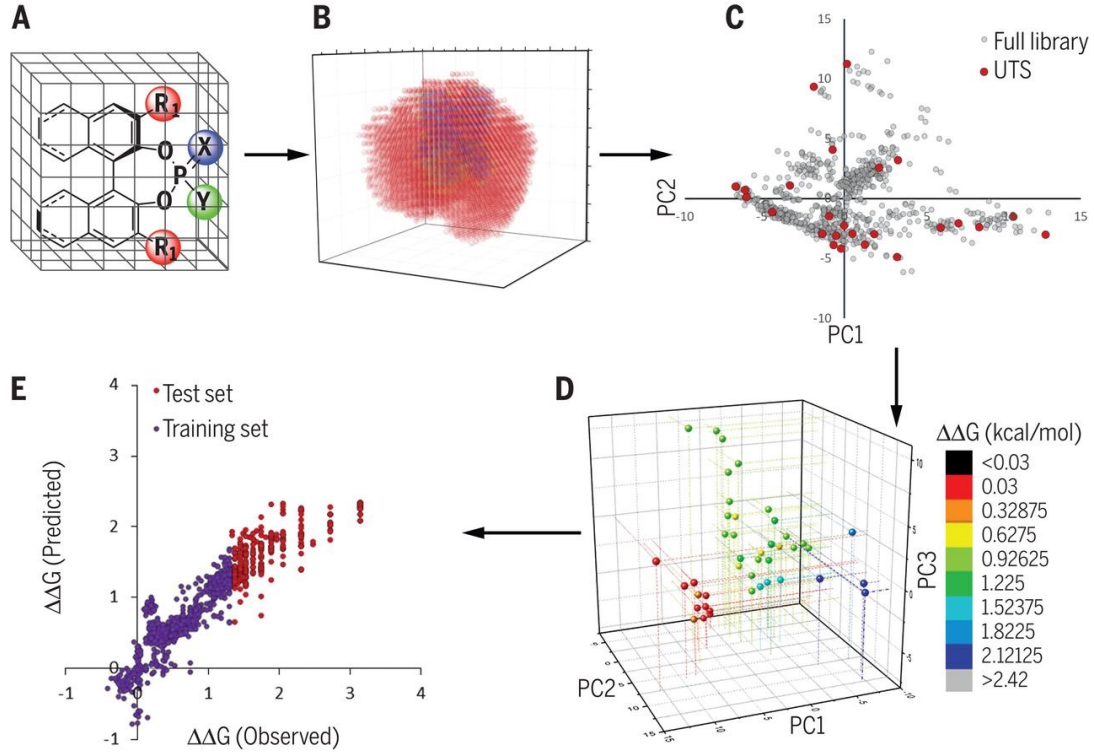


Difference in energy barrier (enantioselectivity) predicted by a linear free energy relationship (LFER)

Structure-based example

Zahrt et al. *Science* 363(6424) **2019**

Zahrt et al. *Chem. Rev.* 120(3), 1620-1689, **2020**

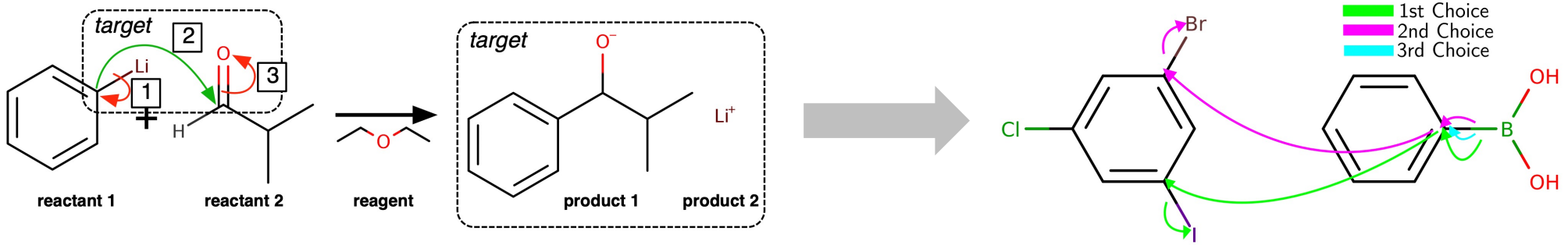


See also: Corminboeuf's NaviCatGA

Mechanistic elucidation from reaction data

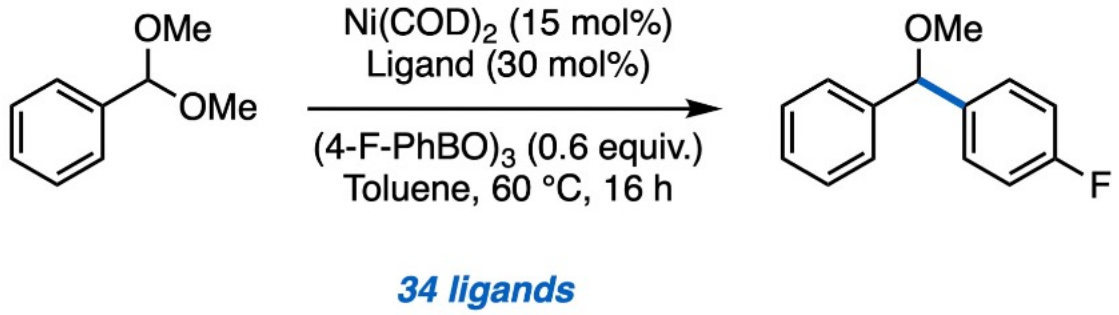
Input: ?; **Output**: mechanistic understanding

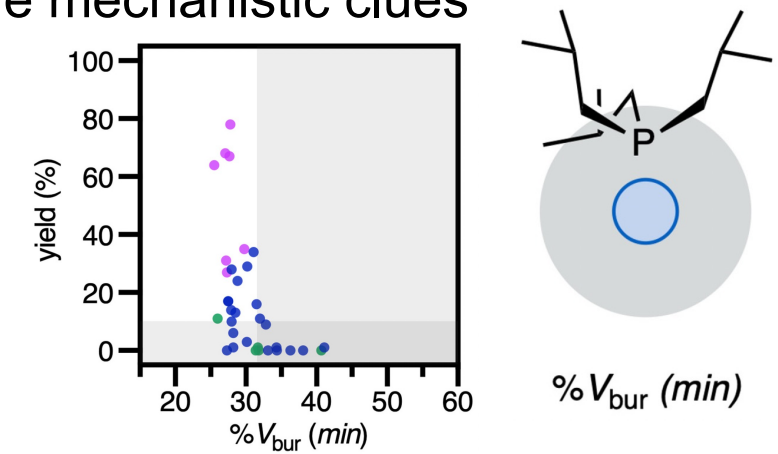
- Not a lot here yet!
- Models trained to think pseudo-mechanistically aren't actually learning mechanisms Bradshaw et al. *ICLR* 2019



- *Post hoc* analysis of learned relationships (e.g., Sigman-esque LFERs) can reveal descriptor importance; in rare cases, univariate relationships can provide mechanistic clues

Newman-Stonebraker et al. *Science* 374(6565) 301-3098, **2021**





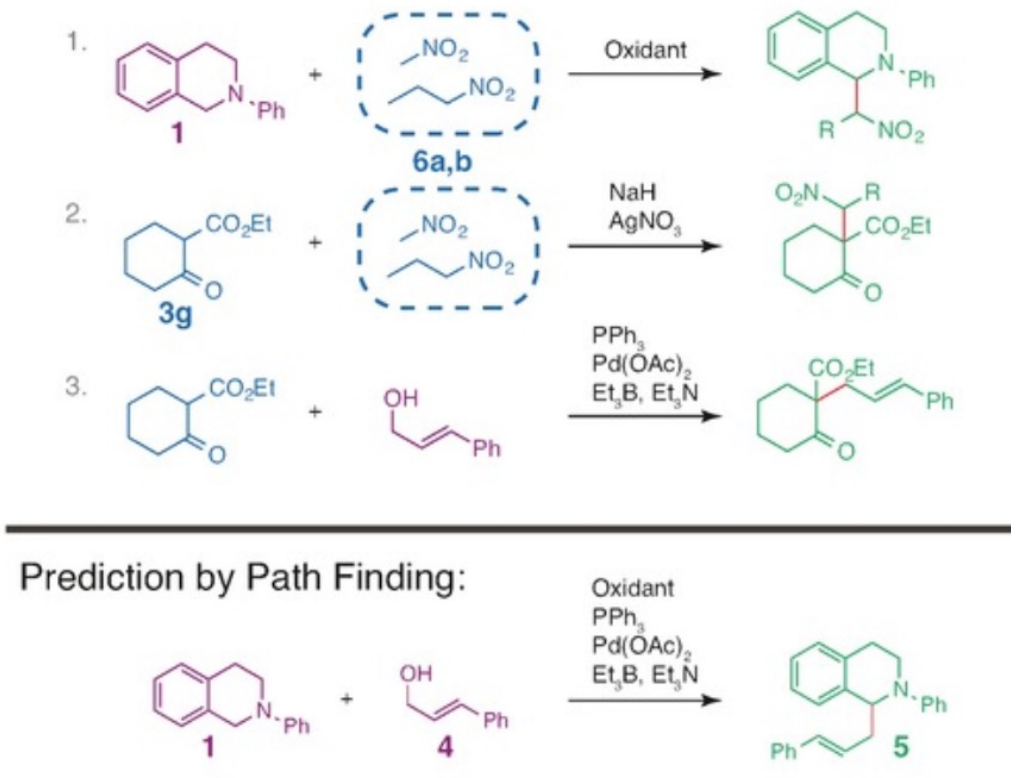
New method development and reaction discovery

Input: ?; **Output**: novel reaction, not just novel substrate

- Also not a lot here – but what is a “new method” or “new reaction” anyway?

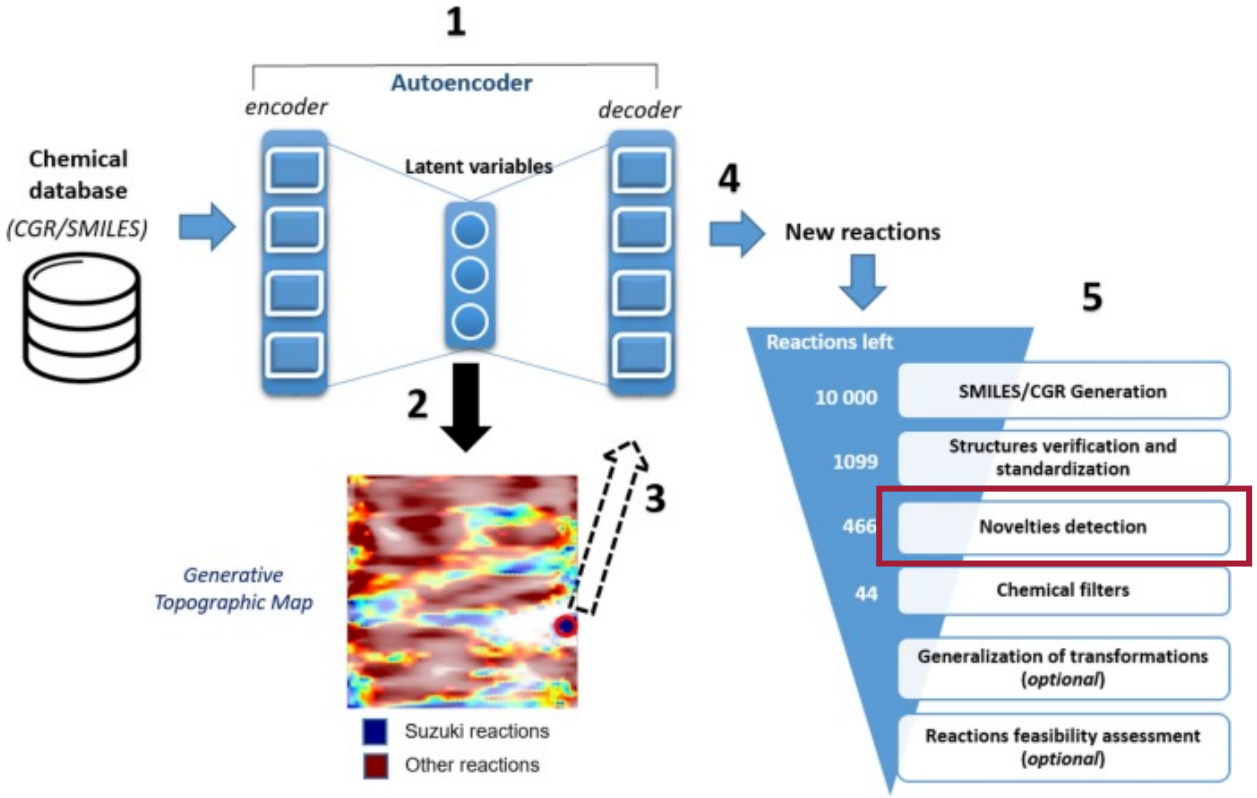
New combinations of known half reactions

Segler and Waller *Chem. Eur. J.* 23(25) 6118-6128, **2017**



Generation of new reaction centers or environments

Bort et al. *Sci. Rep.* 11, 3718, **2021**



Reminder: Predictive chemistry tasks

Primary learning objective for this talk: understand the basics of reaction datasets, representation considerations (beyond molecular representation considerations), *and the common learning tasks*

Deployment

Retrosynthesis, reaction product prediction, classification/mapping

Development

Condition recommendation, condition optimization, scope assessment, catalyst design

Discovery

Mechanistic elucidation, new method development

Investing in the future of predictive chemistry with the Open Reaction Database https://open-reaction-database.org/

Kearnes et al. *JACS* 145(45) 18820-18826, **2021**

- **Building predictive models for chemistry relies on the availability of structured reaction data**
- The ORD is an initiative to "support machine learning and related efforts in reaction prediction, chemical synthesis planning, and experiment design”

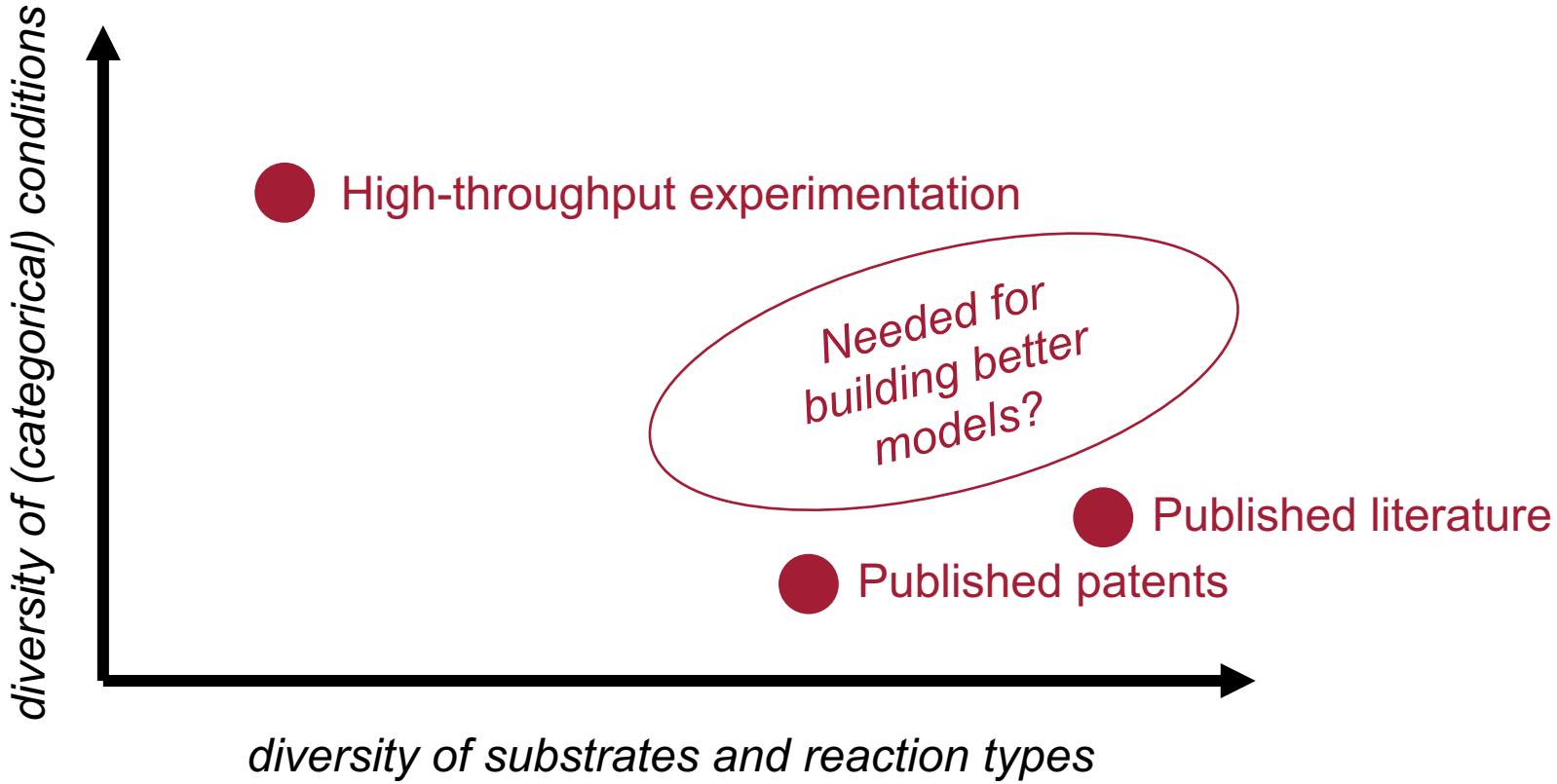
1. Provide a structured data format for chemical reaction data
2. Provide an interface for easy browsing and downloading of data
3. Make reaction data freely and publicly available for anyone to use
4. Encourage sharing of precompetitive proprietary data

Governing Committee
Connor Coley (MIT)
Abby Doyle (UCLA, C-CAS)
Spencer Dreher (Merck)
Joel Hawkins (Pfizer)
Klavs Jensen (MIT)
Steven Kearnes (Relay)

Contributors
Anton Kast (Google)
Brian Lee (Google)
Ian Pendleton (Relay)

Advisory Board
Juan Alvarez (Merck)
Alán Aspuru-Guzik (Toronto, MADNESS)
Tim Cernak (Michigan, Entos)
Lucy Colwell (Cambridge, SynTech, Google)
Werngard Czechtizky (AstraZeneca)
JW Feng (Google)
Matthew Gaunt (Cambridge, SynTech)
Alex Godfrey (NCATS)
Mimi Hii (Imperial, ROAR)
Greg Landrum (T5 Informatics)
Fabio Lima (Novartis)
Christos Nicolaou (Lilly)
Sarah Reisman (Caltech)
Francesco Rianjongdee (GSK)
Matthew Sigman (Utah, C-CAS)
Jay Stevens (BMS)
Sarah Trice (Entos)
Huimin Zhao (UIUC, MMLI)

Information sources affect method development



- Currently absent from databases: order of addition, addition speed, ambient temperature and humidity, reagent purity, chemical vendor, …
  - Missing from literature, and no variation within an HTE dataset
- Diversity of concentrations and reaction times is poor in HTE, even if reagent/catalyst identity varies

Misc. comments on reaction informatics

Chemists in industry do use predictive chemistry tools routinely for route scouting

1. Discovery chemists using routes as proposed, process chemists using for idea generation
2. Data-driven methods can be retrained easily on the most recent reaction data

Data-driven predictive chemistry tools can accelerate chemical development, but they are *not*

1. Providing precise suggestions that are immediately actionable (e.g., using robotics)
2. Expanding synthetically-accessible chemical space by inventing new synthetic methods
3. Removing the need for expert chemist expertise
4. Helping with complex natural product synthesis
5. Perfectly generalizing from very small datasets
6. Operating at the mechanistic level (except Baldi and coworkers)

Acknowledgements

MLPDS

Machine Learning for Pharmaceutical Discovery and Synthesis

AMGEN · BASF We create chemistry · Bristol Myers Squibb · gsk · janssen · Lilly · MERCK · NOVARTIS · Pfizer · SUNOVION · syngenta · 药明康德 WuXi AppTec

Klavs Jensen
William Green
Regina Barzilay
Tommi Jaakkola
Chris Voigt
Daniela Rus
Sangeeta Bhatia
Timothy Jamison

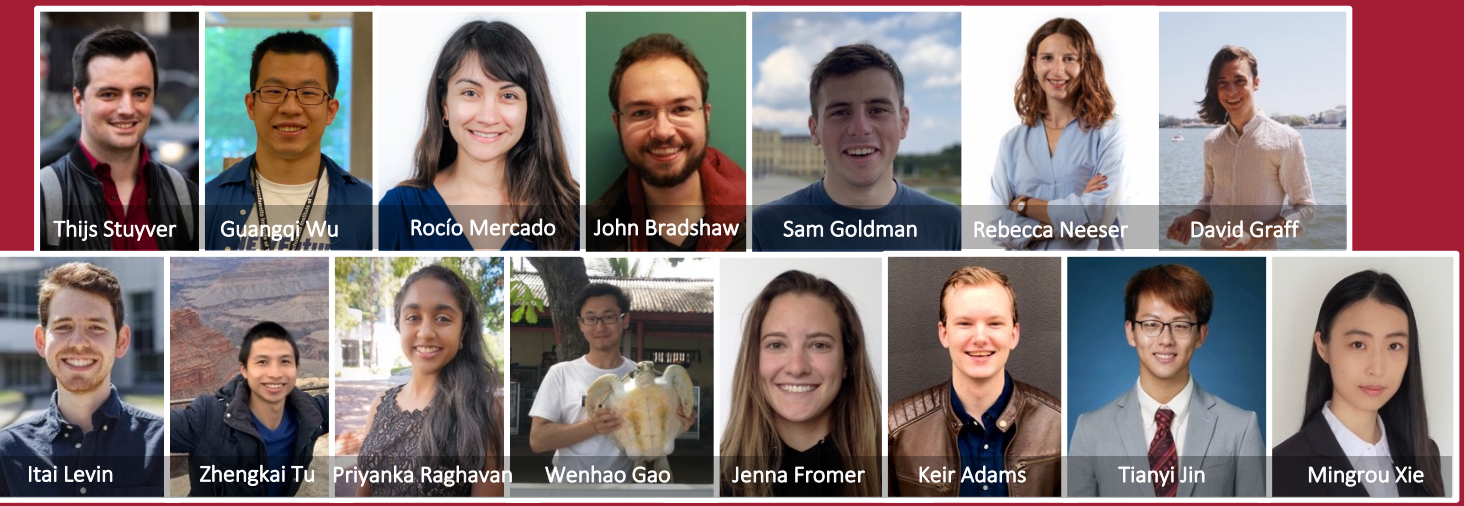
Lucky Pattanaik
Hanyu Gao
Yanfei Guan
Thomas Struble

Vignesh Ram
Ava Soleimany
Alexander Amini
Natalie Eyke

Wengong Jin
Mike Fortunato
Pieter Plehiers
John Schreck
Xiaoxue Wang
Yiming Mo

Dale Thomas
Justin Lummiss
Luke Rogers
Jon Jaworski
Chris Breen
Victor Schultz

Tianfan Fu
Jimeng Sun



+ Saul Vega Sauceda, Janet Li, Divya Nori, Chanwoo Yoon, Ron Shprints

Funding
MLPDS, DARPA, ONR, NIH, NSF, Takeda, Nanite Bio, J-Clinic, MIT-IBM Watson AI Lab

Data
CAS, Pistachio, Reaxys

Advising
Entos, Galixir, Dow, Revela, Kebotix, Anagenex

ccoley@mit.edu | coley.mit.edu | @cwcoley