Modern Approaches to Methods Development

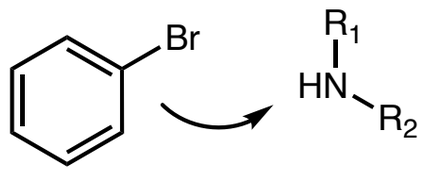
Johnny Wang

Group Meeting

Feb. 14, 2025

The path to developing a useful method

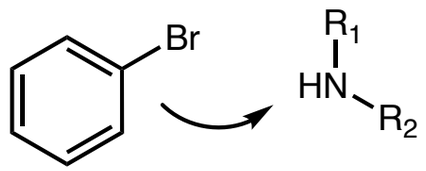
1. Identify a problem



C-N bonds are prevalent but hard to form

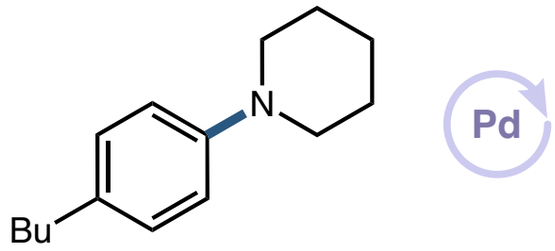
The path to developing a useful method

1. Identify a problem



C-N bonds are prevalent but hard to form

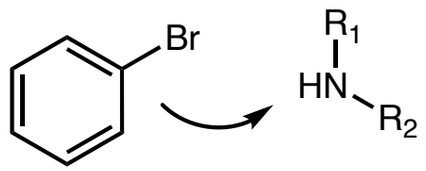
2. Initial hit



1983 - migita

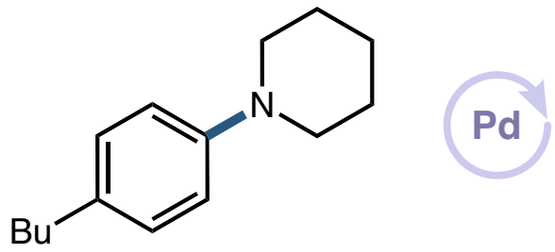
The path to developing a useful method

1. Identify a problem



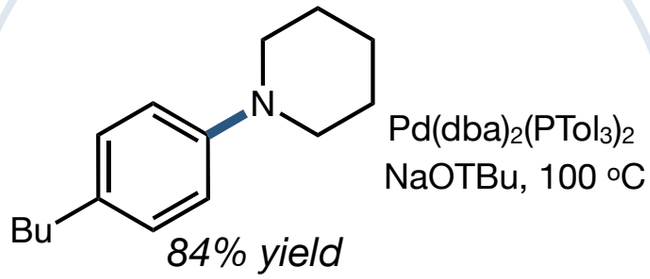
C-N bonds are prevalent but hard to form

2. Initial hit



1983 - migita

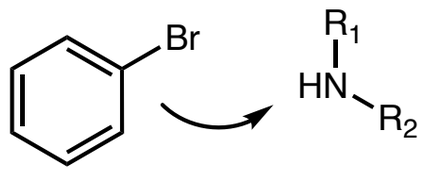
3. Optimization



1994 - Buchwald/Hartwig

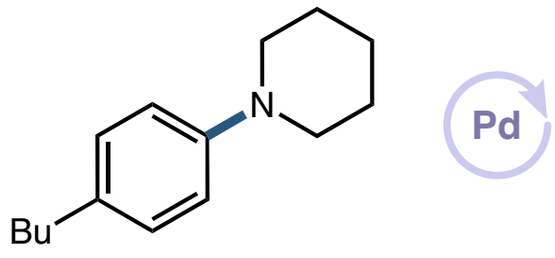
The path to developing a useful method

1. Identify a problem



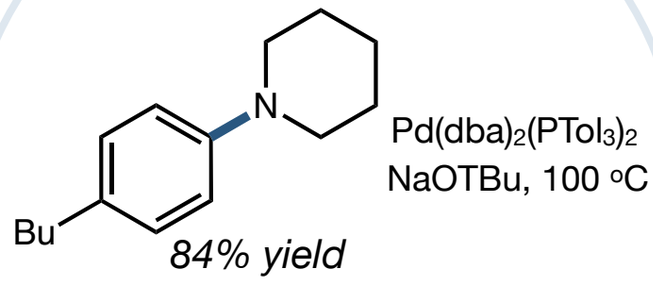
C-N bonds are prevalent but hard to form

2. Initial hit



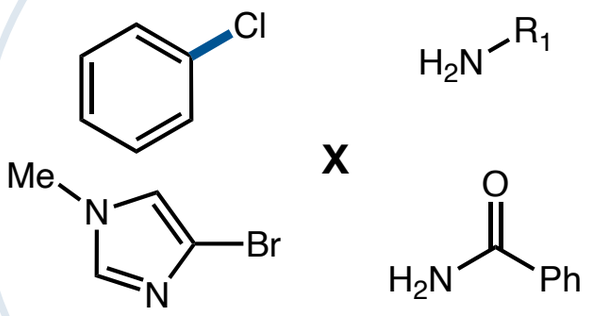
1983 - migita

3. Optimization



1994 - Buchwald/Hartwig

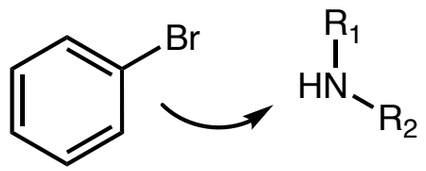
4. Generalize



~ongoing (but mostly done)

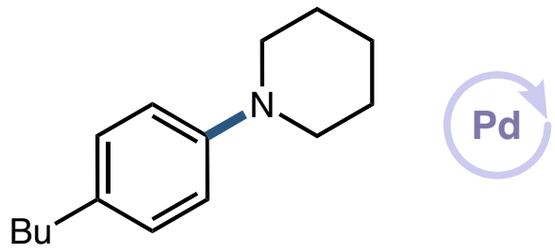
The path to developing a useful method

1. Identify a problem



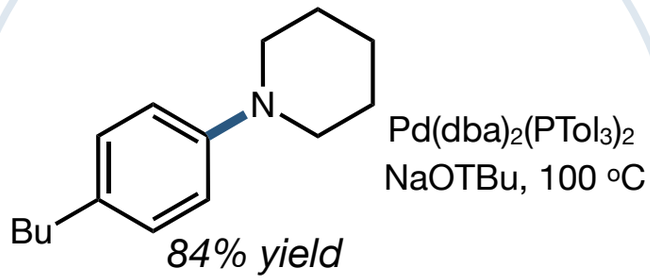
C-N bonds are prevalent but hard to form

2. Initial hit



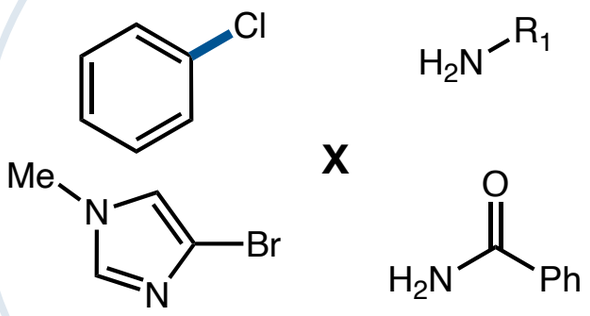
1983 - migita

3. Optimization



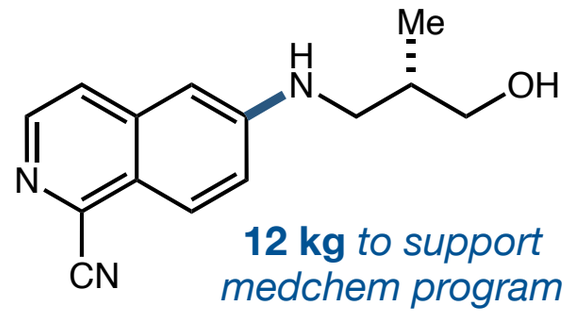
1994 - Buchwald/Hartwig

4. Generalize



~ongoing (but mostly done)

5. General adoption



Widely adopted

Needs to be highly predictable

Traditional approach to methods development

Initial discovery 1983

Today

- *Thousands of publication*
- *100's of ligands*
- *100's of substrate combinations*
- *Among most used reactions in industry*

OPRD **2019**, *23*, 9.; *J. Med. Chem.* **2016**, *59*, 4443.; *Chem. Sci.* **2020**, *11*, 13085.

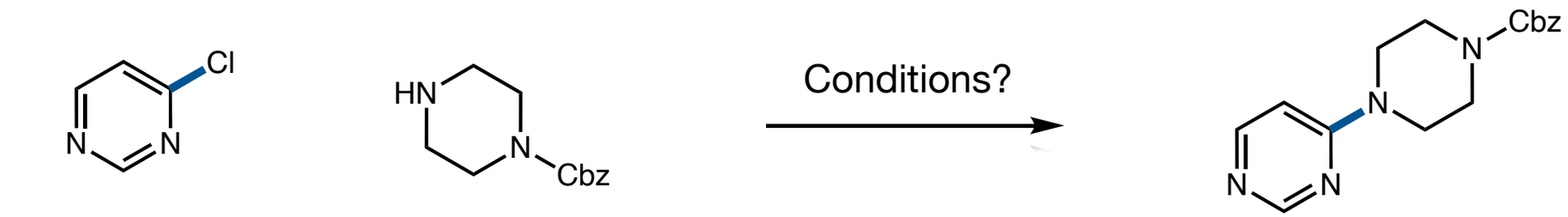
Traditional approach to methods development

Initial discovery 1983

Today

- *Thousands of publication*
- *100's of ligands*
- *100's of substrate combinations*
- *Among most used reactions in industry*

Conditions and ligands known for almost any given combination



heteroaryl chloride *dialkyl amine*

OPRD **2019**, *23*, 9.; *J. Med. Chem.* **2016**, *59*, 4443.; *Chem. Sci.* **2020**, *11*, 13085.

Traditional approach to methods development

Initial discovery 1983

Today

- *Thousands of publication*
- *100's of ligands*
- *100's of substrate combinations*
- *Among most used reactions in industry*

Electrophile Type	Aryl	DiAryl	Alkyl	DiAlkyl	Alkyl-Aryl	aromN	Ketimine	Amide
I_HAR	BINAP, Cs_2CO_3 78% Xantphos, Cs_2CO_3 60% BINAP, K_2CO_3 54%			Xantphos, NaOtBu 61% BINAP, NaOtBu 52% Xantphos, Cs_2CO_3 36%				Xantphos, Cs_2CO_3 74%
I_ARY	DPEphos, NaOtBu 88% SPhos, NaOtBu 74% dppf, NaOtBu 74%	Xantphos, NaOtBu 81% $P(tBu)_3$, KOtBu 77% $P(tBu)_3$, NaOtBu 72%		BINAP, NaOtBu 59% $P(o\text{-}tol)_3$, NaOtBu 58% BINAP, Cs_2CO_3 57%		$P(Ph)_3$, NaOtBu 80% $P(tBu)_3$, NaOtBu 72%		Xantphos, Cs_2CO_3 54%
Cl_HAR	XPhos, KOtBu 90% 142691-72-3, NaOtBu 76% CyJohnPhos, NaOtBu 70%	$P(tBu)_3$, NaOtBu 70%	BrettPhos, LiHMDS 78% BINAP, K_2CO_3 73% Cy-tBu-Josiphos, NaOtBu 68%	DavePhos, KOtBu 80% Triisobutylphosphatrane, NaOtBu 80% DavePhos, NaOtBu 76%	BINAP, Cs_2CO_3 87% DavePhos, K_2CO_3 84% Xantphos, Cs_2CO_3 40%	$P(tBu)_3$, KOtBu 80% $P(tBu)_3$, K3PO4 76% $P(tBu)_3$, NaOtBu 70%	BINAP, Cs_2CO_3 62%	dCypf, Cs_2CO_3 96% dppf, K_3PO_4 85% BINAP, Cs_2CO_3 60%
Cl_ARY	SIPr, KOtBu 98% BrettPhos, NaOtBu 94% $P(Ph)_3$, NaOtBu 88%	$P(tBu)_3$, NaOtBu 75% SPhos, NaOtBu 60%	Cy-tBu-Josiphos, NaOtBu 94% Ad-BippyPhos, KOPh 90% Cy-tBu-Josiphos, LiHMDS 85%	Triisobutylphosphatrane, NaOtBu 83% XPhos, NaOtBu 71% JohnPhos, NaOtBu 60%		cBRIDP, NaOtBu 78% $P(tBu)_3$, NaOtBu 68%		JohnPhos, Cs_2CO_3 94% tBuBrettPhos, K_3PO_4 92% BrettPhos, Cs_2CO_3 51%
Br_HAR	DPEphos, KOtBu 83% BINAP, KOtBu 79% tBuXPhos, NaOtBu 70%	$P(tBu)_3$, NaOtBu 68% dppf, NaOtBu 60%	Cy-tBu-Josiphos, NaOtBu 72% BINAP, NaOtBu 62% Xantphos, NaOtBu 52%	Xantphos, NaOtBu 73% RuPhos, LiHMDS 71% RuPhos, Cs_2CO_3 70%		Xantphos, Cs_2CO_3 72% $P(tBu)_3$, NaOtBu 65% $P(Ph)_3$, NaOtBu 64%	BINAP, Cs_2CO_3 74% BINAP, NaOtBu 66%	BINAP, Cs_2CO_3 72% Xantphos, Cs_2CO_3 54% XPhos, Cs_2CO_3 46%
Br_ARY	SPhos, Cs_2CO_3 95% QUINAP, NaOtBu 91% QUINAP, Cs_2CO_3 90%	RuPhos, LiHMDS 94% $P(tBu)_3$, NoBase 81% $P(tBu)_3$, KOtBu 79%	Ad-BippyPhos, KOPh 95% Cy-tBu-Josiphos, NaOtBu 94% Xantphos, KOtBu 86%	XPhos, NaOtPent 83% Triisobutylphosphatrane, NaOtBu 78% RuPhos, LiHMDS 77%	$P(tBu)_3$, NaOtBu 76% JohnPhos, NaOtBu 75% BINAP, Cs_2CO_3 74%	tBuXPhos, NaOtBu 90% $P(tBu)_3$, K_3PO_4 83% $P(o\text{-}tol)_3$, NaOtBu 80%	dppf, NaOtBu 88% BINAP, NaOtBu 85% Xantphos, Cs_2CO_3 64%	dppf, NaOtBu 81% cBRIDP, NaOtBu 80% Xantphos, K_3PO_4 80%

OPRD **2019**, *23*, 9.; *J. Med. Chem.* **2016**, *59*, 4443.; *Chem. Sci.* **2020**, *11*, 13085.

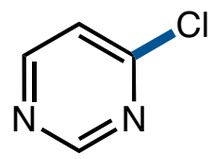
Traditional approach to methods development

Initial discovery 1994

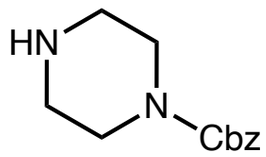
Today

- *Thousands of publication*
- *100's of ligands*
- *100's of substrate combinations*
- *Among most used reactions in industry*

Conditions and ligands known for almost any given combination

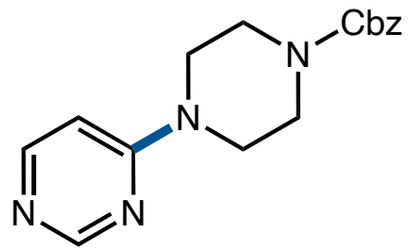


heteroaryl chloride



dialkyl amine

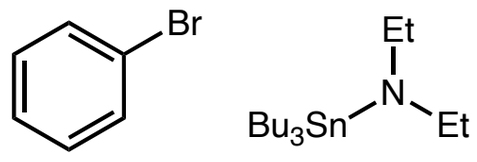
BrettPhos, LiHMDS	78%
BINAP, K_2CO_3	73%
Cy-tBu-Josiphos, NaOtBu	68%



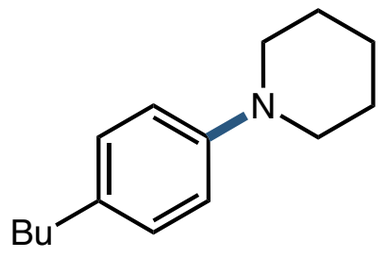
OPRD **2019**, *23*, 9.; *J. Med. Chem.* **2016**, *59*, 4443.; *Chem. Sci.* **2020**, *11*, 13085.

Traditional approach to methods development-belaboring the point

Initial hit -1983



First practical system -1994



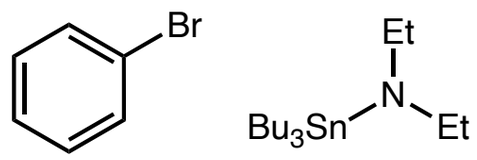
Today: *Mostly generalized*

40 years, 1390 publications

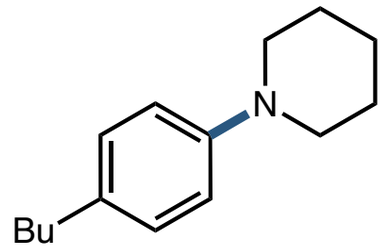
New combinations still being optimized

Traditional approach to methods development-belaboring the point

Initial hit -1983



First practical system -1994

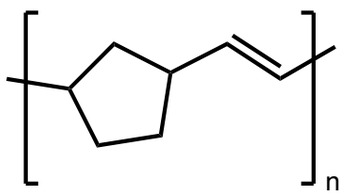


Today: *Mostly generalized*

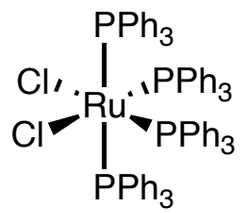
40 years, 1390 publications

New combinations still being optimized

Initial hit -1955



First practical system -1992



Today: *Mostly generalized*

70 years, 15,622 publications

New combinations still being optimized

Development of a general, useful reaction is slow

Traditional approach to methods development

How can the “generalization” and development of new reactions be greatly accelerated?

Modern approaches to methods development

Modern Paradigms in Screening

- *Reaction generalization*
- *“Accelerate” Serendipity*
- *Miniaturization of unique reaction set ups*

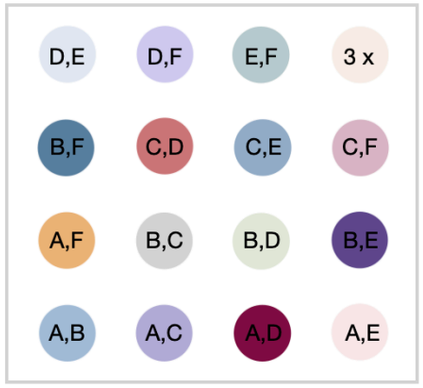
Data Science

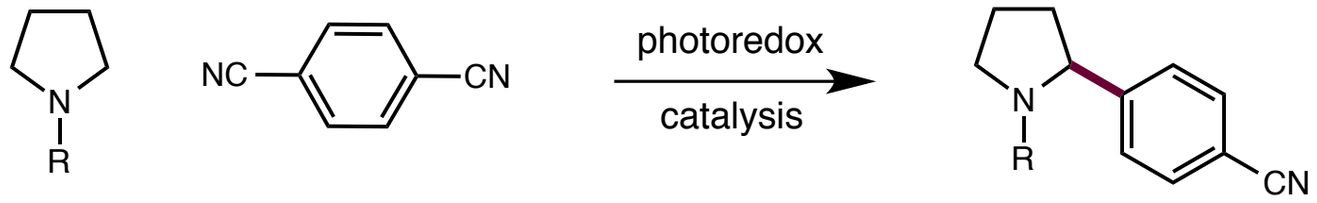
- *Catalyst optimization*
- *Predicting selectivity*
- *Discovery of new catalysts*

Machine Learning

- *What is machine learning*
- *Prediction of optimal conditions*
- *Selectivity prediction for complex systems*
- *Catalyst Discovery*

Accelerated Serendipity



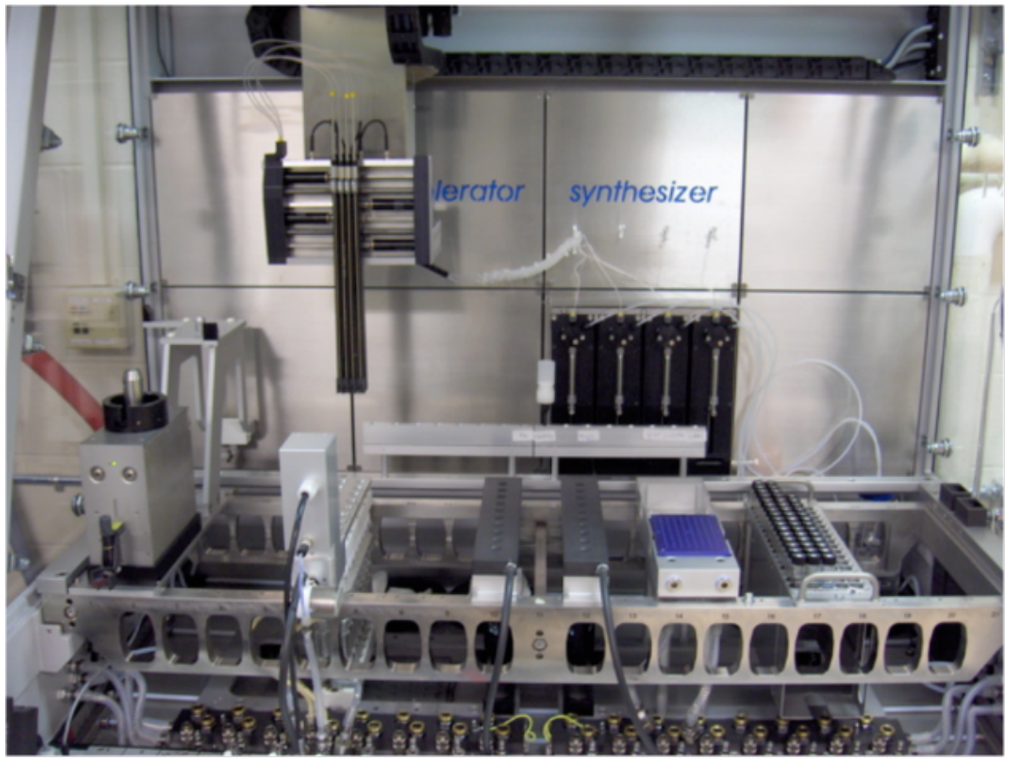


Discovery of new amine C-H functionalization through accelerated serendipity

Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

Accelerated Serendipity

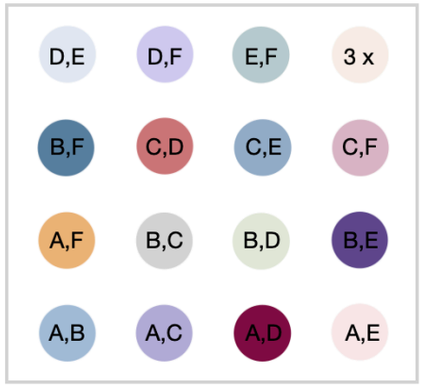
Puts together 500 - 1,000 reactions per day in a highly controlled fashion

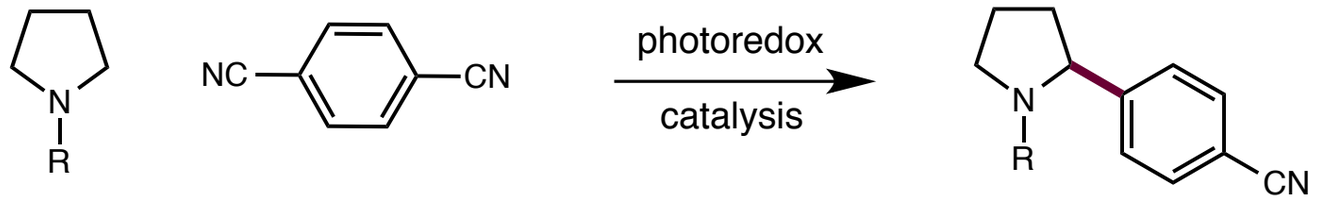


- Parallel processing of reactions at multiple temperatures in multiple solvents
- Multi channel filtration and vacuum capabilities, solid weighing and work ups performed
- 96 well LED plates with easy installation of LEDs of variable wavelengths

Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

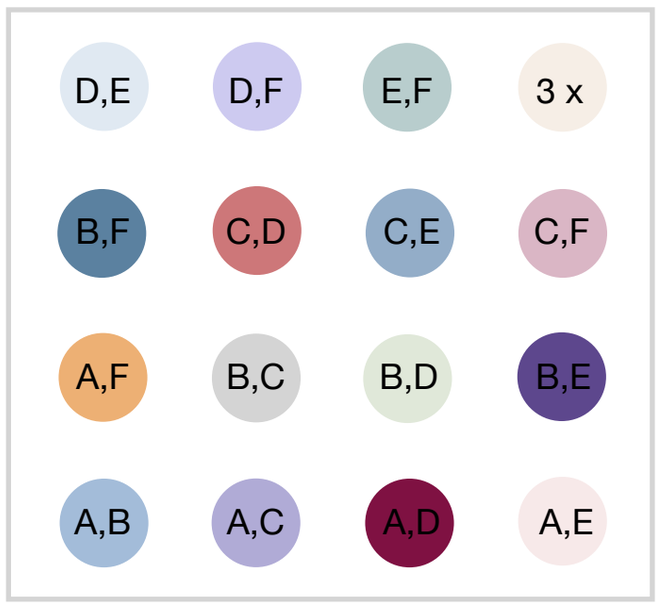
Accelerated Serendipity





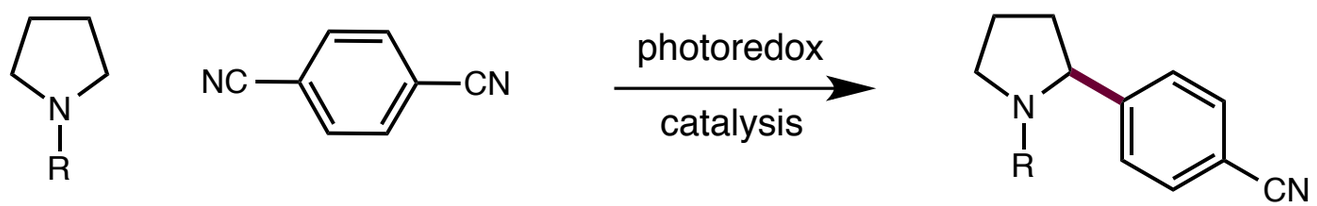
Discovery of new amine C-H functionalization through accelerated serendipity

19 x 19 matrix



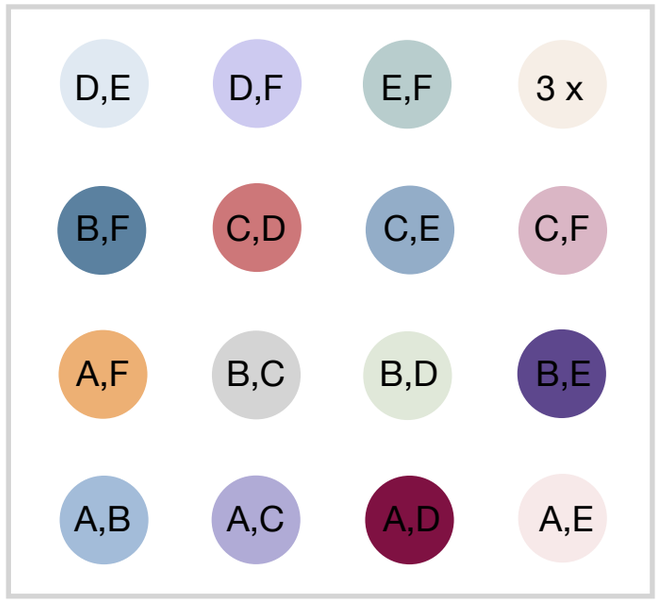
Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

Accelerated Serendipity



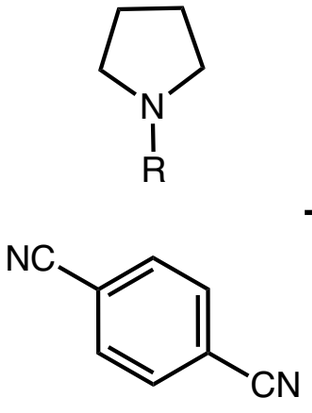
Discovery of new amine C-H functionalization through accelerated serendipity

19 x 19 matrix



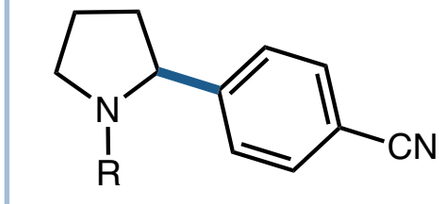
Reaction Evaluation

$Ru(bpy)_3Cl_2$
0.5 mole%



Re-test

Initial Result

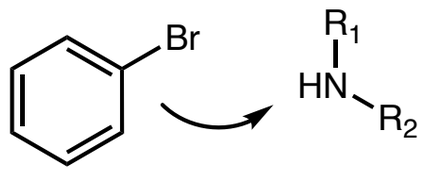


7% yield

Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

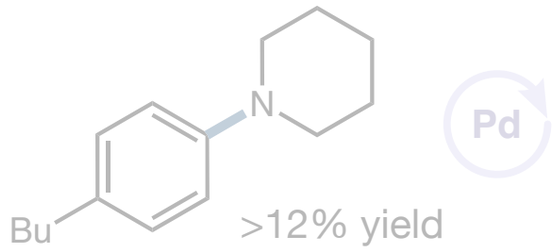
Accelerated Serendipity

1. Identify a problem



C-N bonds are prevalent but hard to form

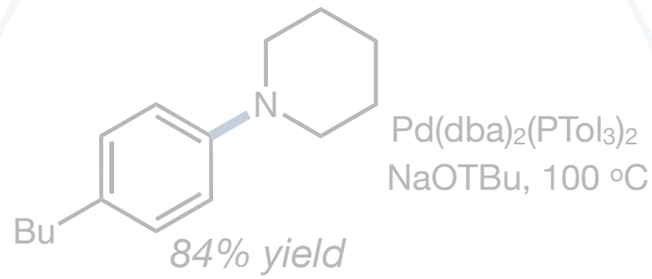
2. Initial hit



>12% yield

Bond formation at low yields

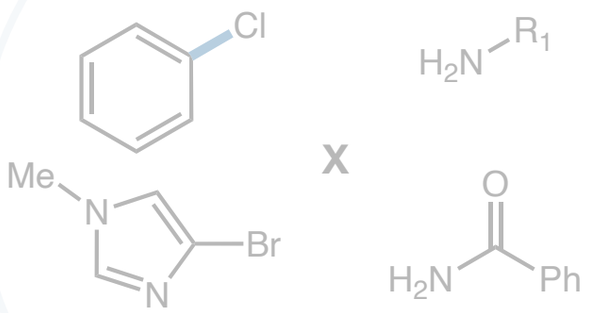
3. Optimization



84% yield

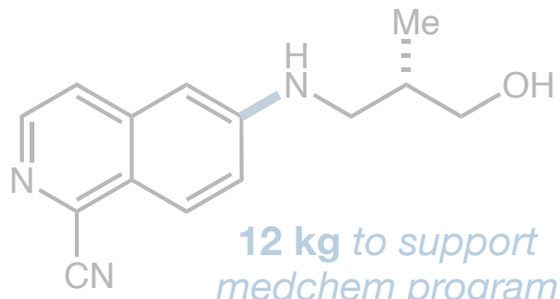
Conditions give high yields for initial system

4. Generalize



Works for most substrate combinations

5. General adoption



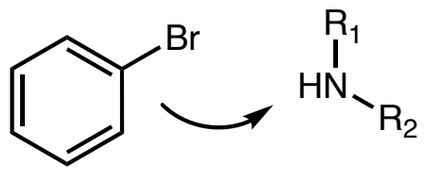
12 kg *to support medchem program*

Wide adoption by community

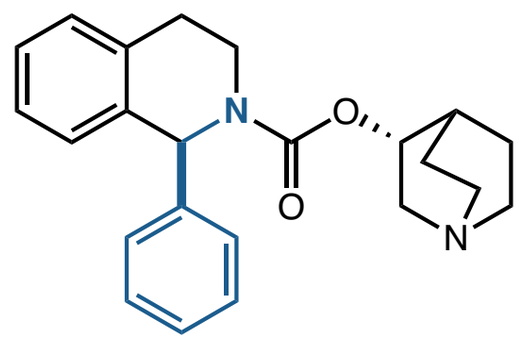
Needs to have easily predictable results

Accelerated Serendipity

1. Identify a problem

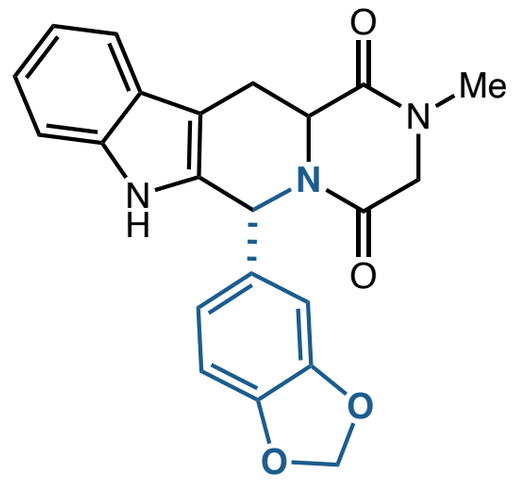


C-N bonds are prevalent but hard to form



Vesicare (GSK) No. 115/200

Muscarinic receptor antagonist



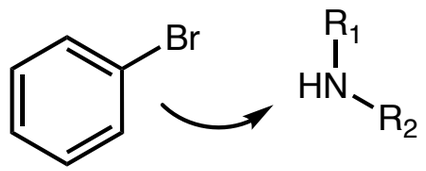
Cialis (Lilly) No. 66/200

Phosphodiesterase inhibitor

Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

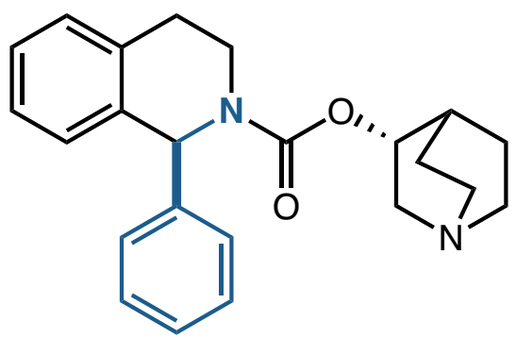
Accelerated Serendipity

1. Identify a problem



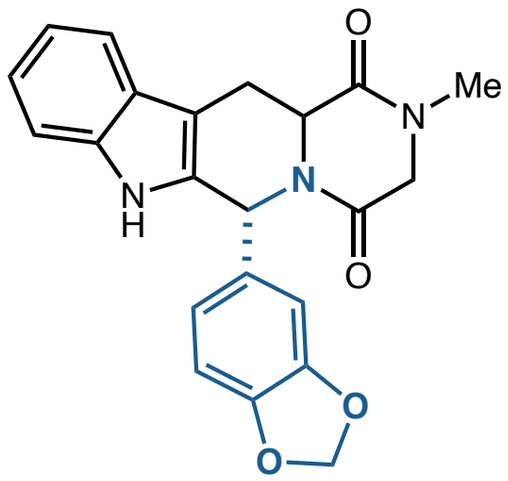
C-N bonds are prevalent but hard to form

C-H arylation is an extremely desirable reaction



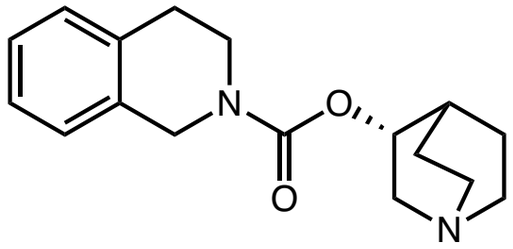
Vesicare (GSK) No. 115/200

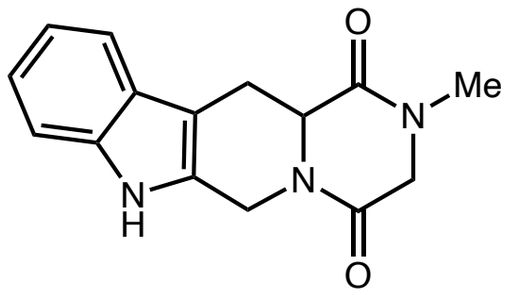
Muscarinic receptor antagonist



Cialis (Lilly) No. 66/200

Phosphodiesterase inhibitor

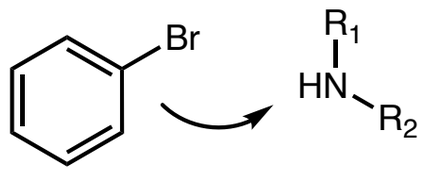




Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

Accelerated Serendipity

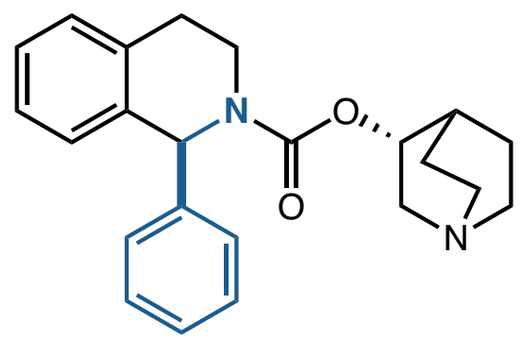
1. Identify a problem



C-N bonds are prevalent but hard to form

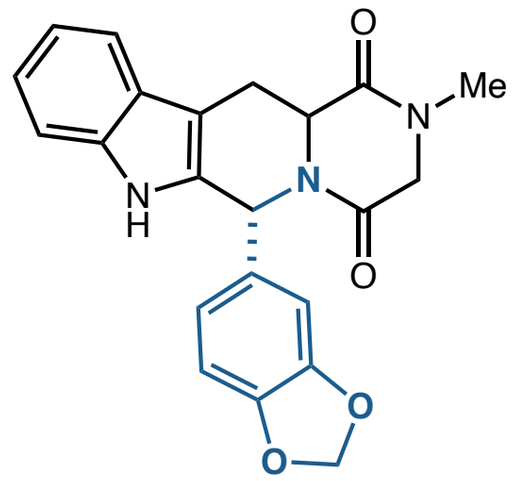
C-H arylation is an extremely desirable reaction

Not compatible with the synthetic logic of the time!



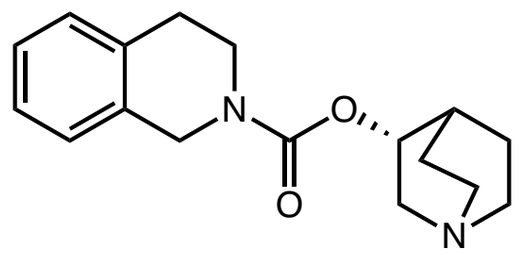
Vesicare (GSK) No. 115/200

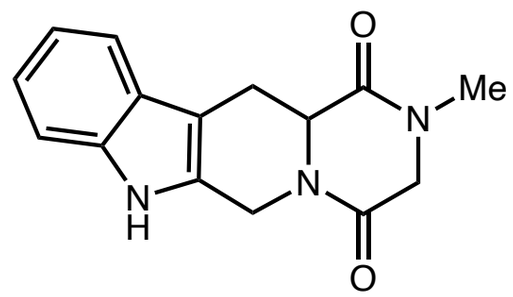
Muscarinic receptor antagonist



Cialis (Lilly) No. 66/200

Phosphodiesterase inhibitor

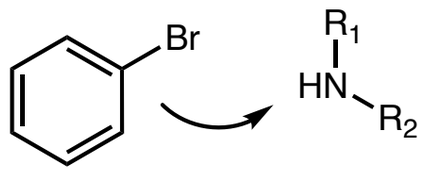




Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

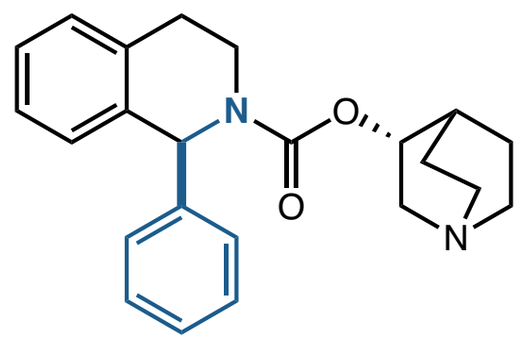
Accelerated Serendipity

1. Identify a problem



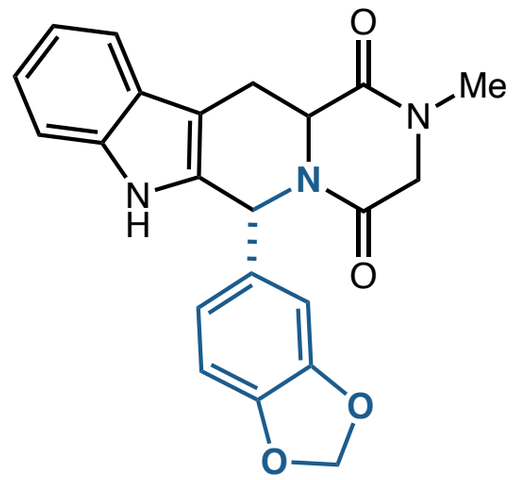
C-N bonds are prevalent but hard to form

“Predict” a problem and solve it at the same time



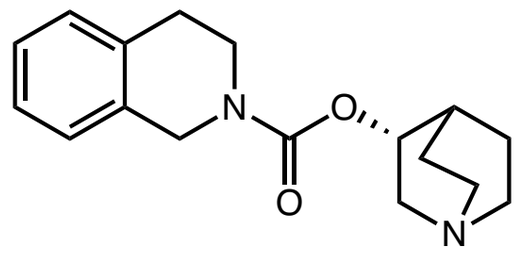
Vesicare (GSK) No. 115/200

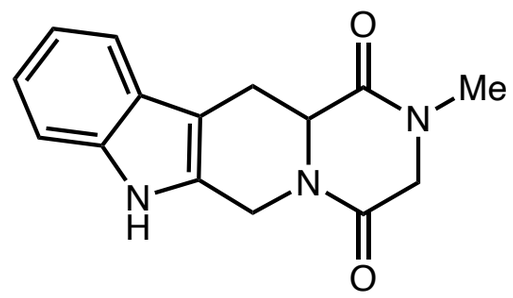
Muscarinic receptor antagonist



Cialis (Lilly) No. 66/200

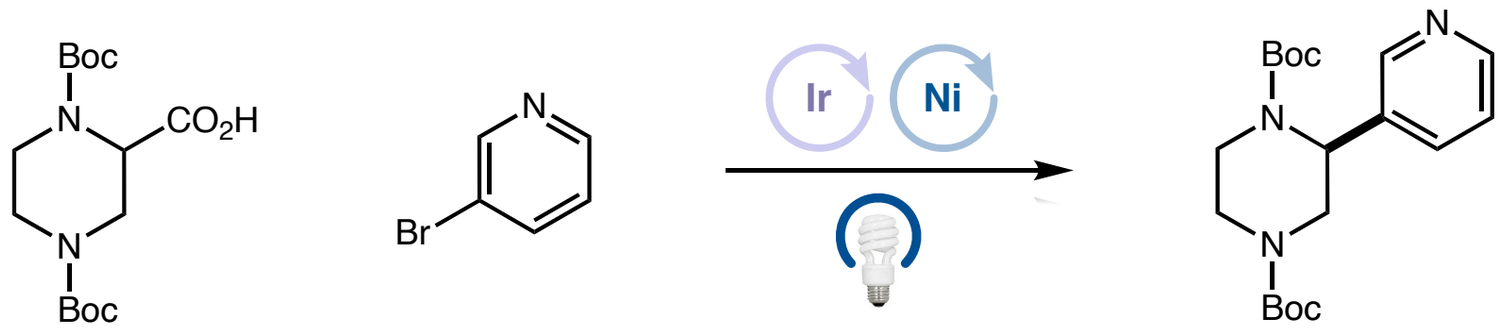
Phosphodiesterase inhibitor





Mcnally, A.; Prier, C. K.; MacMillan, D. W. C.; *Science* **2011**, *334*, 1114–1117

Additive Screening



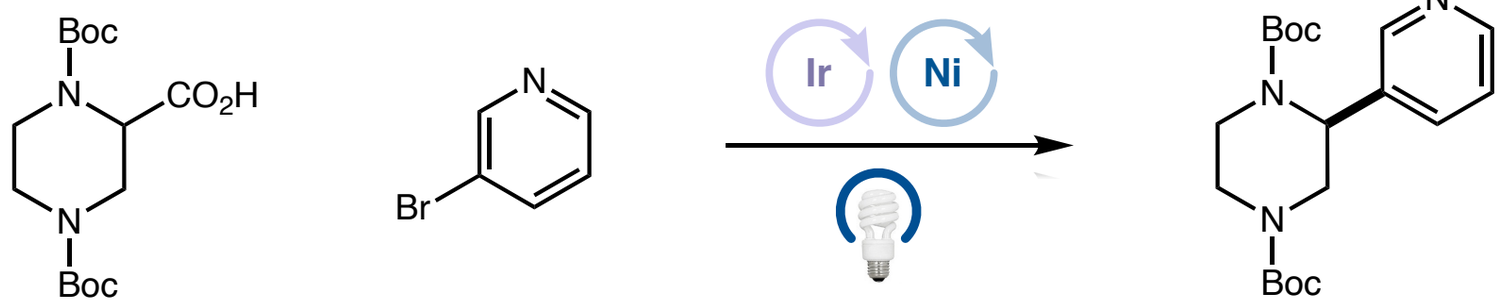
Can this reaction be generalized faster than historical approaches?

Design a new catalyst for every problematic combination

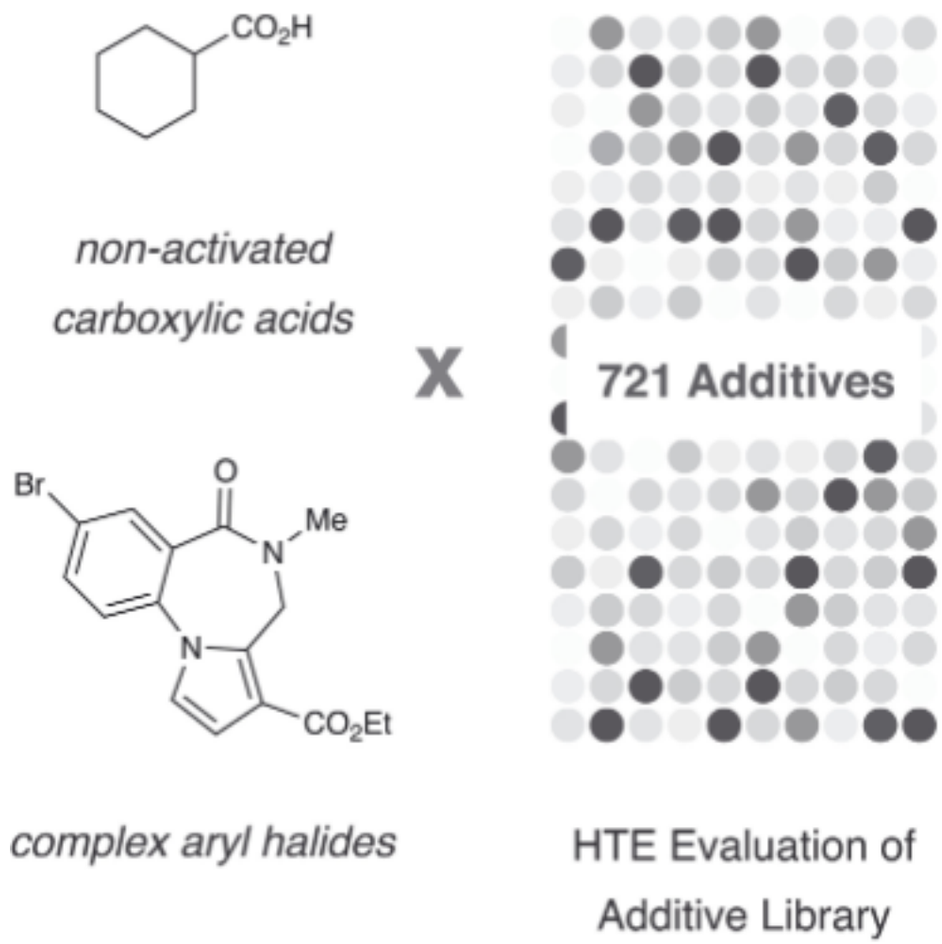
Universal additive that helps all substrate combinations?

Kullmer, C. N. P.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D., MacMillan, D. W. C. *Science* **2022**, *376*, 532–539

Additive Screening



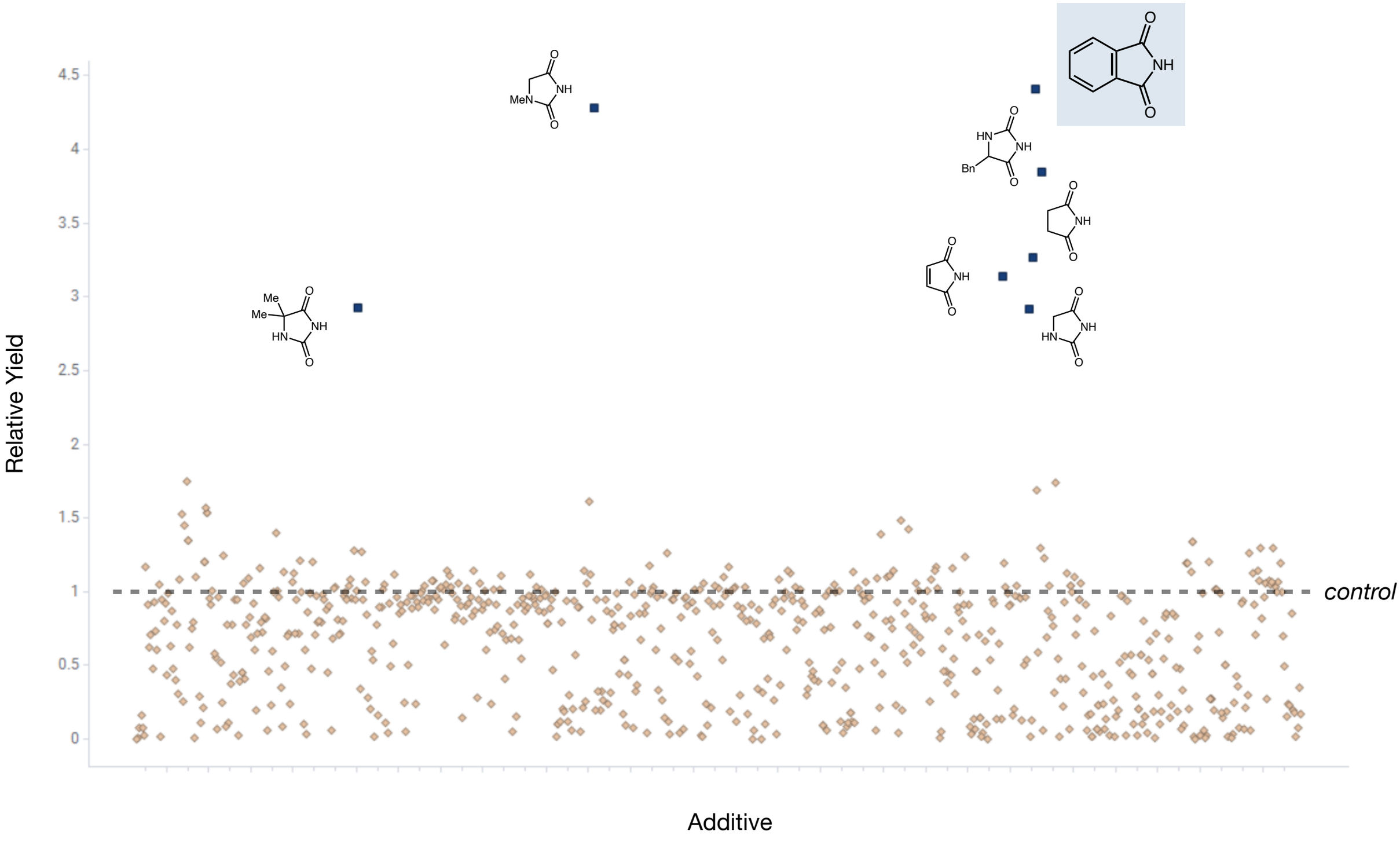
Can this reaction be generalized faster than historical approaches?



Universal additive that helps all substrate combinations?

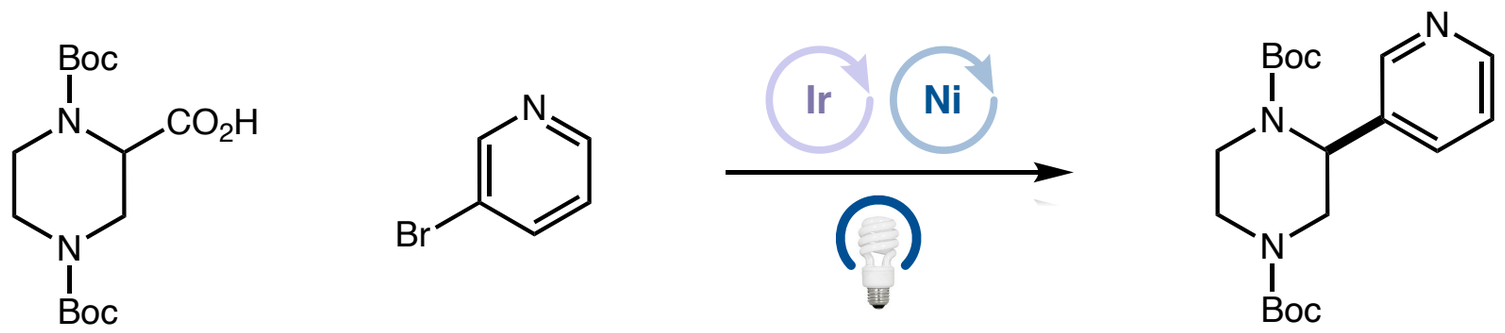
Kullmer, C. N. P.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D., MacMillan, D. W. C. *Science* **2022**, *376*, 532–539

Additive Screening

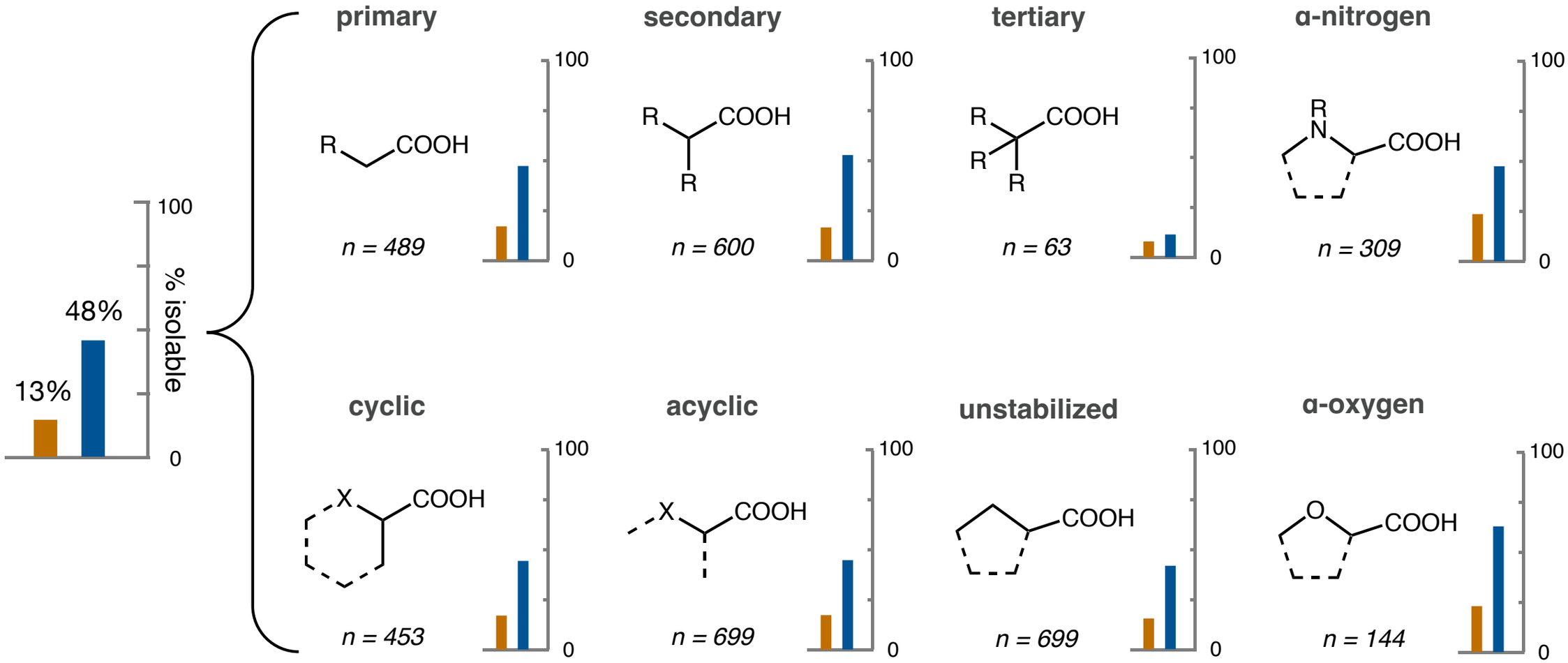


Kullmer, C. N. P.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D., MacMillan, D. W. C. *Science* **2022**, *376*, 532–539

Additive Screening

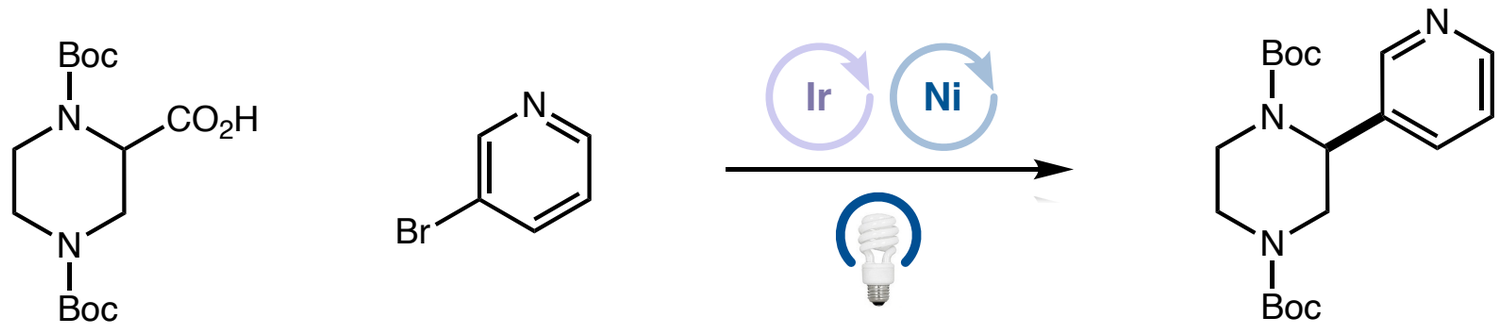


Can this reaction be generalized faster than historical approaches?

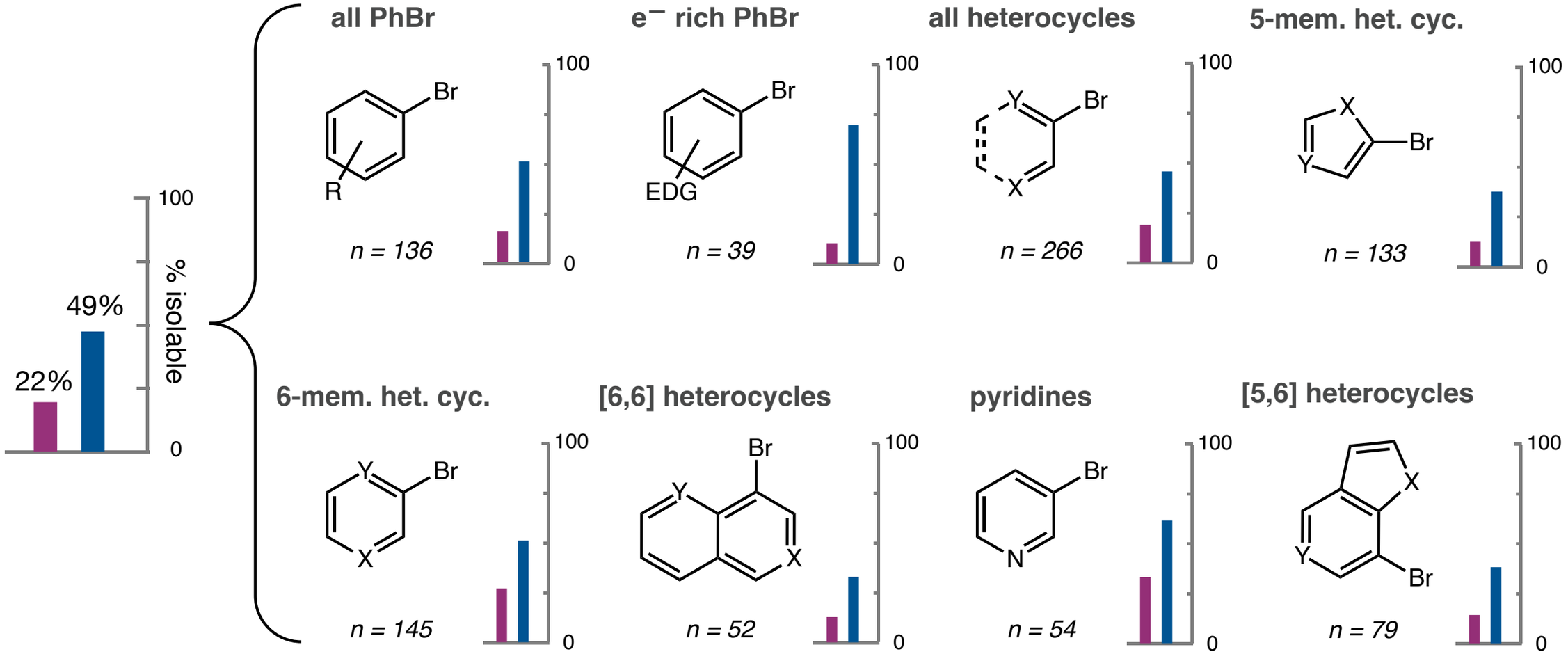


Kullmer, C. N. P.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D., MacMillan, D. W. C. *Science* **2022**, *376*, 532–539

Additive Screening

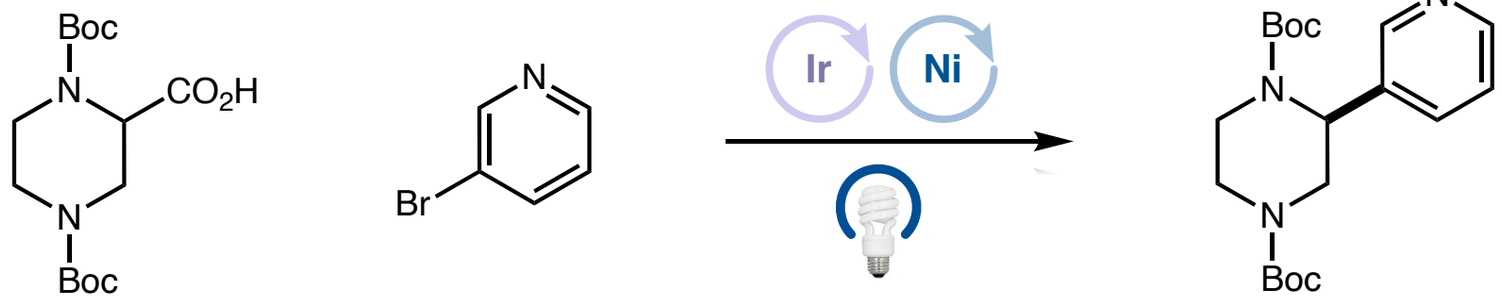


Can this reaction be generalized faster than historical approaches?



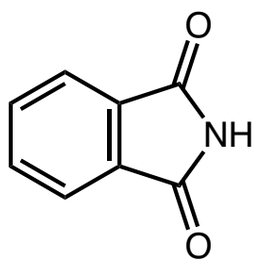
Kullmer, C. N. P.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D., MacMillan, D. W. C. *Science* **2022**, *376*, 532–539

Additive Screening



Can this reaction be generalized faster than historical approaches?

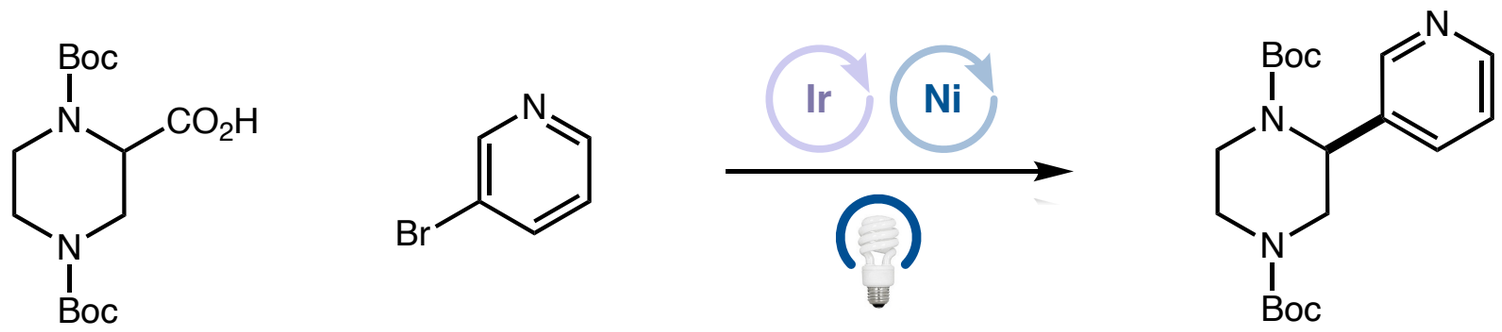
just add phthalimide



Greatly more general reaction

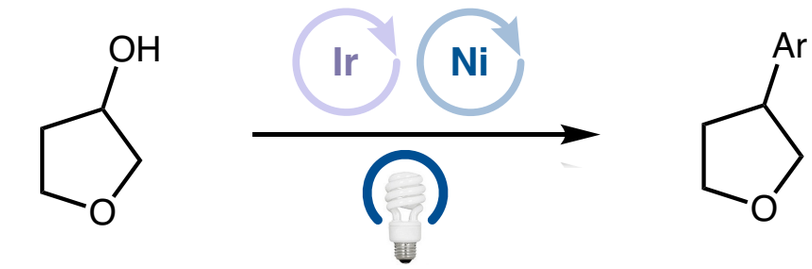
Kullmer, C. N. P.; Kautzky, J. A.; Krska, S. W.; Nowak, T.; Dreher, S. D., MacMillan, D. W. C. *Science* **2022**, *376*, 532–539

Additive Screening



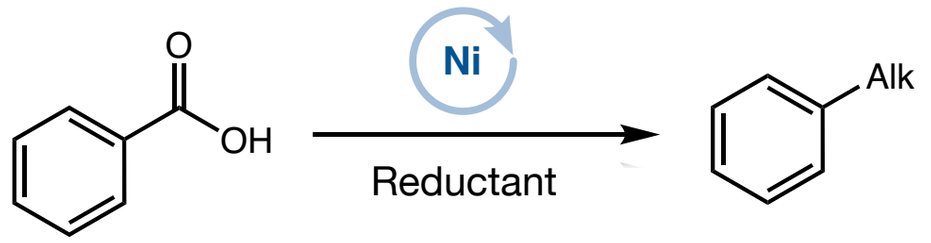
Can this reaction be generalized faster than historical approaches?

Deoxygenative arylation



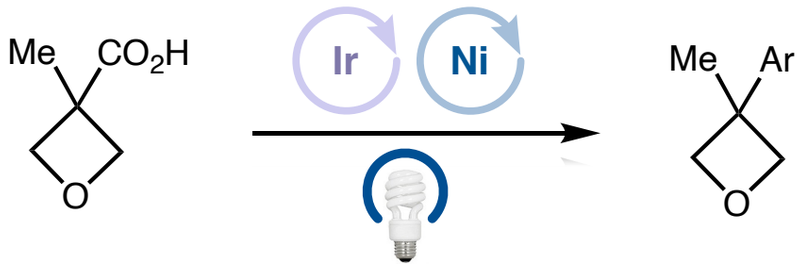
Nature **2021**, *598*, 451–456

Decarboxylative alkylation



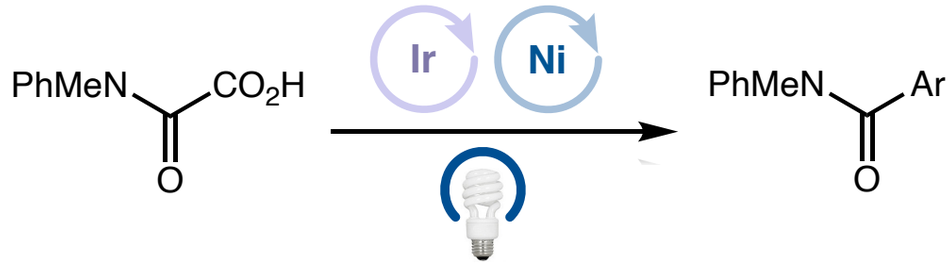
JACS **2023**, *145*, 10930–10937

Tertiary arylation



ACS Catal. **2023**, *13*, 11910–11918

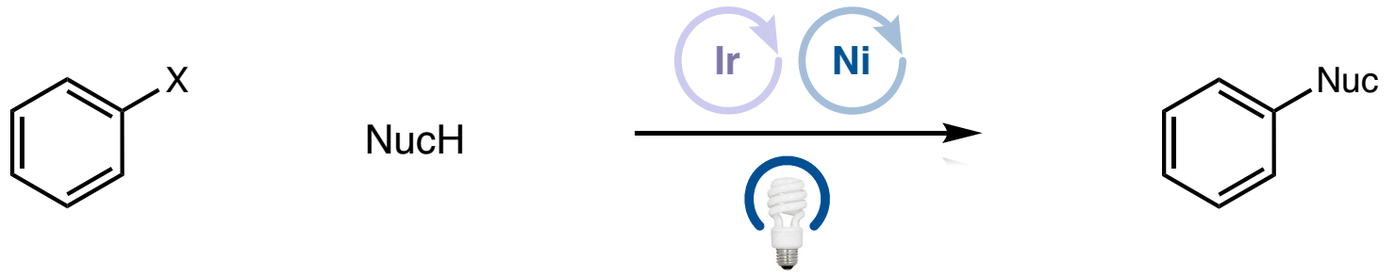
Aminocarbonylation



Org. Lett. **2024**, 26, 2768–2772

Some of many examples where phthalamde has helped with optimizing and generalizing novel methods

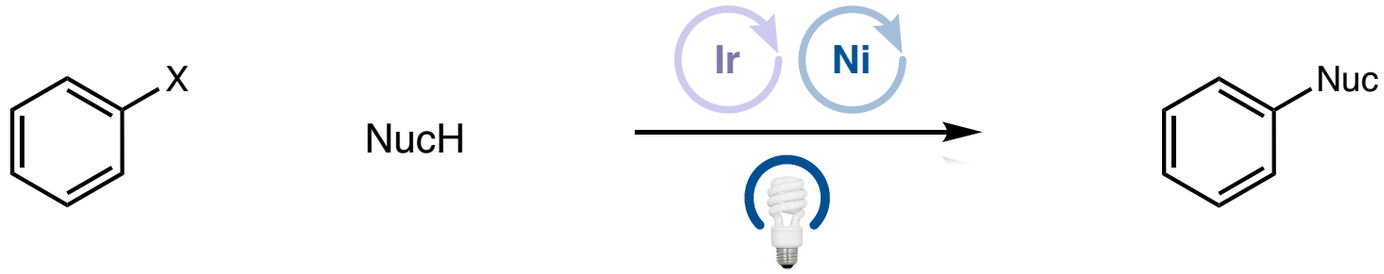
Additive Screening



General conditions using different additives for different Nuc

Ghosh, I.; Shlapakov, N.; Karl, T. A.; Duker, J.; Nikitin, M.; Burykina, J. V.; Ananikov, V. P.; Konig, B. *Nature* **2023**, *619*, 87–93

Additive Screening



General conditions using different additives for different Nuc

Thiols – aliphatic
Thiols – benzylic
Thiophenols – electron rich
Thiophenol
Thiophenols – electron poor
Selenophenol
Diphenylphosphine
Diphenylphosphine oxide
Tetramethylguanidine
N-Boc-guanidine
Benzamidine
Amines – secondary
Amines – primary
Primary amines – (distant functional groups)
Amines – benzylic
Benzophenone imine
Sulfoximines
Pyrazole
Imidazole
Indazole
Dialkylphosphine
H_2O
Anions
Cyanate$^-$
Aryl sulfinates
Alkyl sulfinates
Chloride$^-$
Azide$^-$
Cbz-protected hydrazine
Phenols – electron rich
Anilines – electron rich
Aniline – *p*-Cl
Ethyl phenylphosphinate
Diethylphosphite
B_2pin_2
Anilines – electron poor
Carbazole
Aryl sulfonamides
Alkyl sulfonamides
Benzamide
Aliphatic amides
Cyclic amides
Indole
Benzimidazole
Phenols – electron poor
Aliphatic alcohols
α-Amino carbon
Phenyl hydrazine
Phenylacetylene
Diphenylsilane

Group I
None

Group II

Group II′

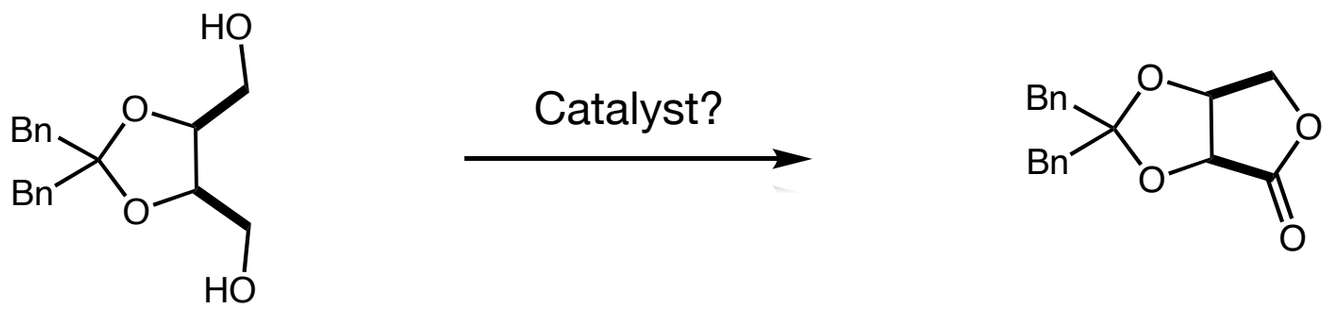
Group III

Group IV

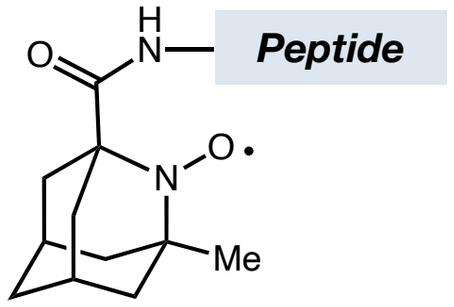
Ghosh, I.; Shlapakov, N.; Karl, T. A.; Duker, J.; Nikitin, M.; Burykina, J. V.; Ananikov, V. P.; Konig, B. *Nature* **2023**, *619*, 87–93

Generalization Through Multi-Substrate Optimization in Enantioselective Catalysis

Enantioselective oxidation



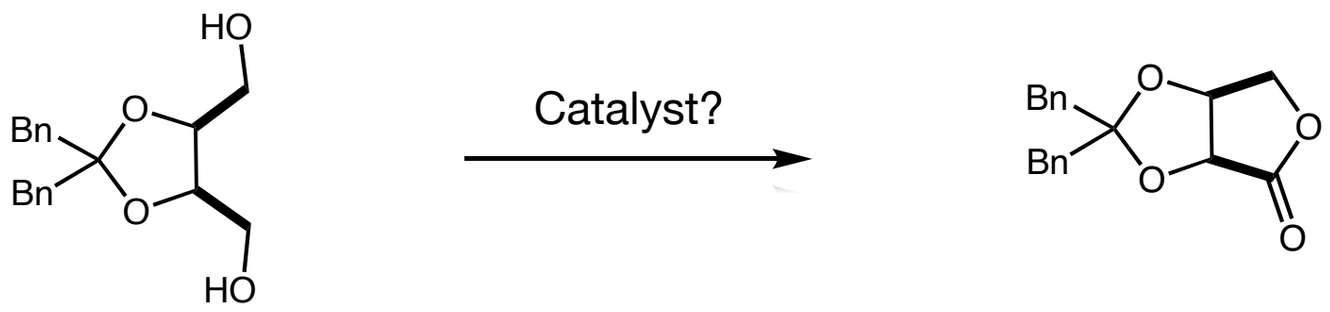
Catalyst Design



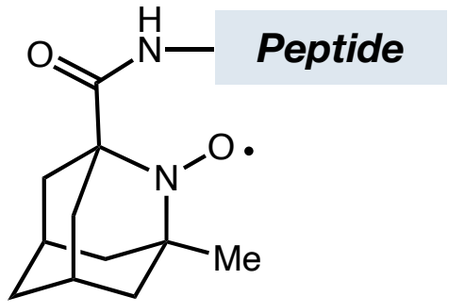
Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. *Science* **2023**, *380*, 701–712

Generalization Through Multi-Substrate Optimization in Enantioselective Catalysis

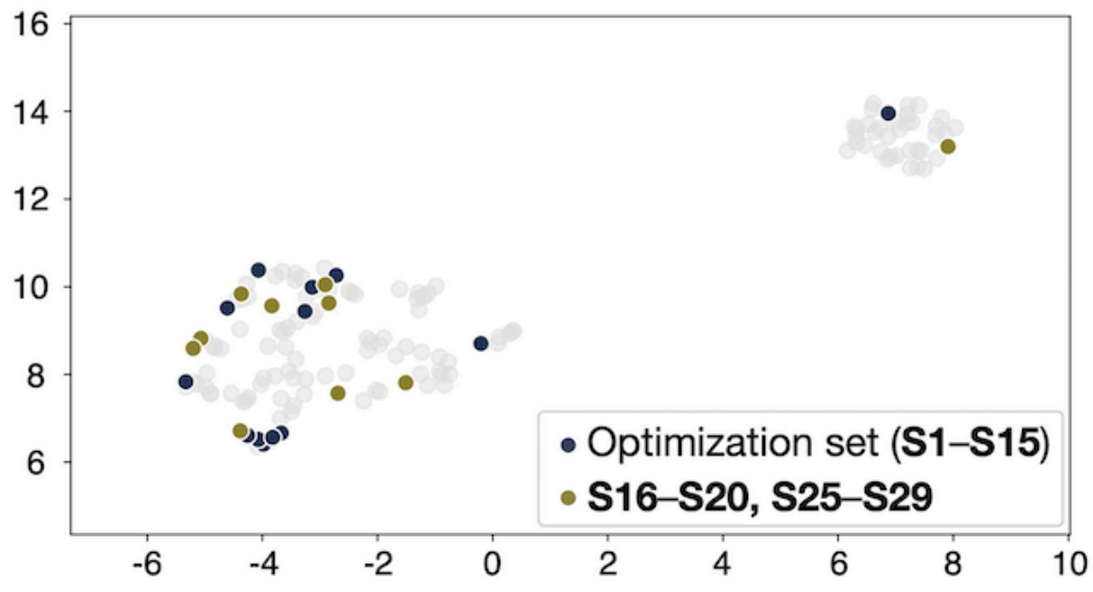
Enantioselective oxidation



Catalyst Design



2D visualization of 1,4 diol chemical space



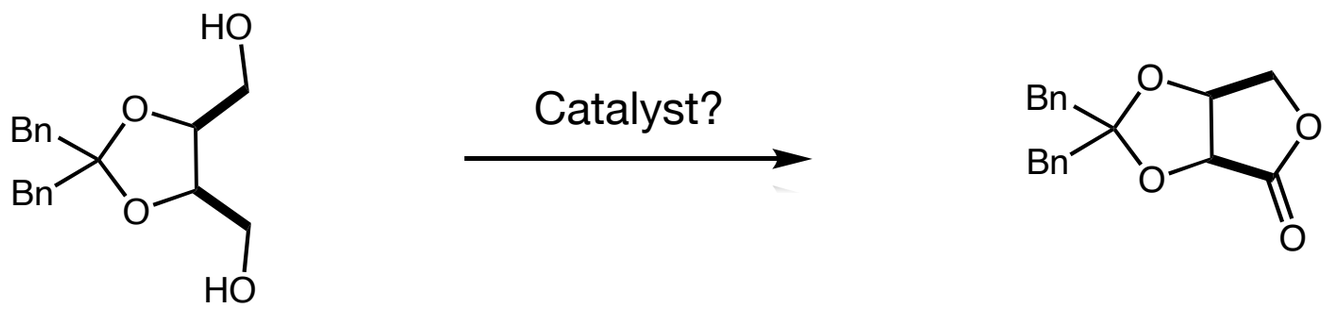
For more details about scope parameterization:

Kariofillis, S. K.; Jiang, S.; Zuranski, A .M.; Gandhi, S. S.; Alvarado, J. I. M.; Doyle, A. G. *J. Am. Chem. Soc.* **2022**, *144*, 1045–≠1055

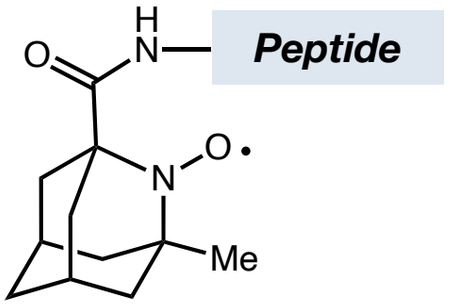
Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. *Science* **2023**, *380*, 701–712

Generalization Through Multi-Substrate Optimization in Enantioselective Catalysis

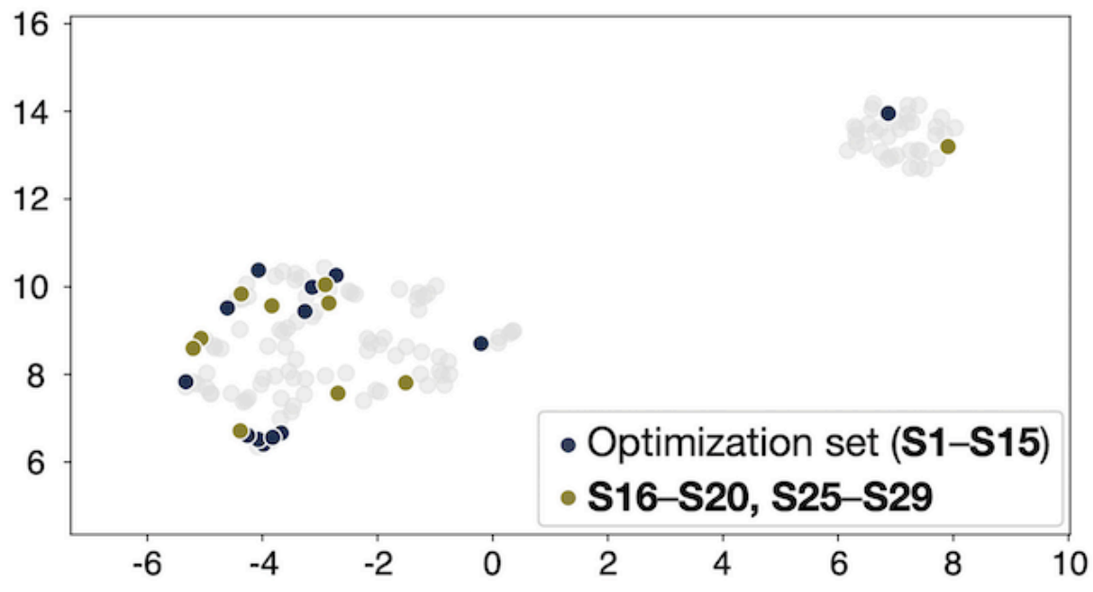
Enantioselective oxidation



Catalyst Design



2D visualization of 1,4 diol chemical space



15 optimization substrates

X

7 rounds of catalyst design

Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. *Science* **2023**, *380*, 701–712

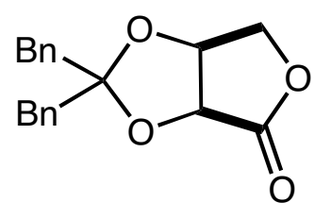
Generalization Through Multi-Substrate Optimization in Enantioselective Catalysis

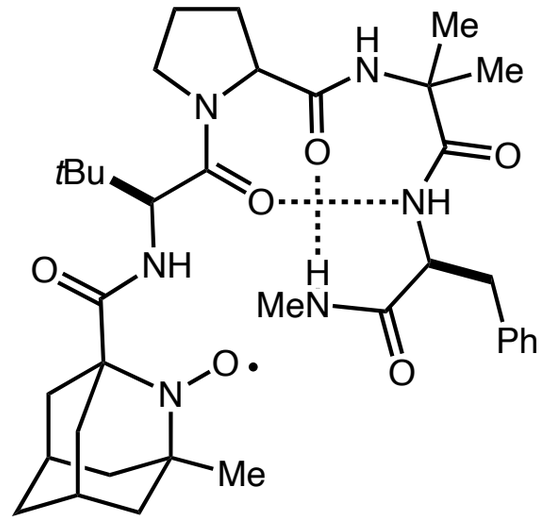
Catalyst

	%ee	P1	P2	P3	P4
Substrate	**S1**	75	75	85	94

Under a one substrate optimization mode P4 is an "optimal catalyst

S1





P4

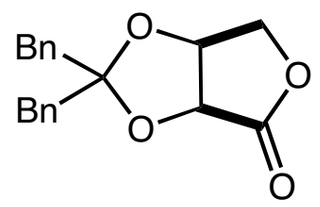
Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. *Science* **2023**, *380*, 701–712

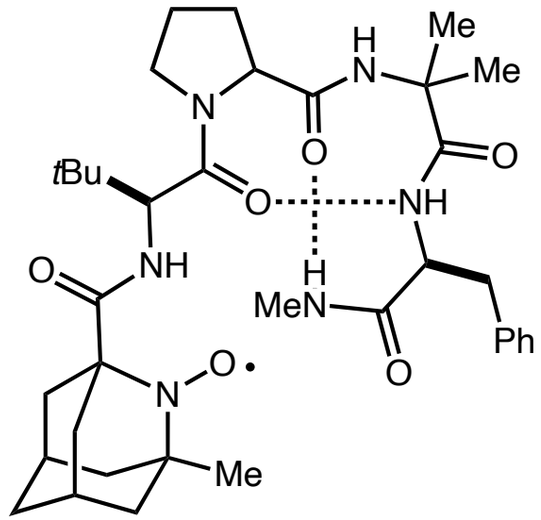
Generalization Through Multi-Substrate Optimization in Enantioselective Catalysis

Catalyst

Substrate %ee	P1	P2	P3	P4
S1	75	75	85	94
S2	6	2	11	69
S3	19	30	26	75
S4	23	33	1	77
S5	68	72	76	86
S6	36	37	69	87
S7	27	40	4	63
S8	2	13	14	71
S9	8	0	29	69
S10	27	40	28	48
S11	3	2	8	46
S12	7	5	17	22
S13	13	29	6	4
S14	9	34	15	1
S15	16	31	9	1
$\%ee_{med}$	16	31	15	69

S1





P4

Not an optimal catalyst in a multi-substrate optimization paradigm

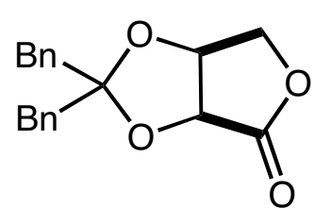
Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. *Science* **2023**, *380*, 701–712

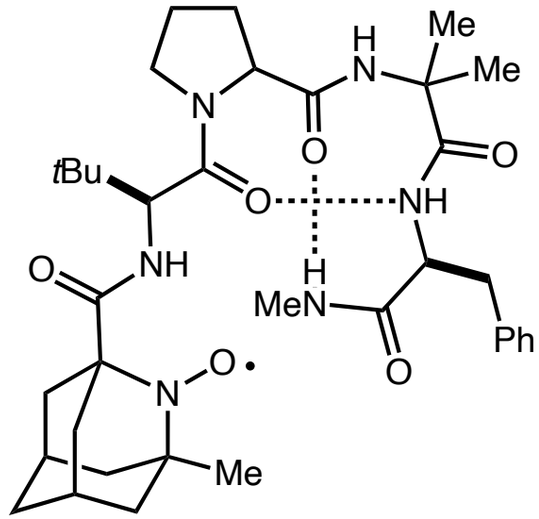
Generalization Through Multi-Substrate Optimization in Enantioselective Catalysis

Catalyst

Substrate / %ee	P1	P2	P3	P4	P5	P6	P7
S1	75	75	85	94	95	95	97
S2	6	2	11	69	83	86	97
S3	19	30	26	75	72	93	96
S4	23	33	1	77	91	90	96
S5	68	72	76	86	86	95	95
S6	36	37	69	87	85	96	94
S7	27	40	4	63	82	83	93
S8	2	13	14	71	74	89	93
S9	8	0	29	69	48	79	93
S10	27	40	28	48	81	68	90
S11	3	2	8	46	36	76	79
S12	7	5	17	22	46	51	79
S13	13	29	6	4	15	40	65
S14	9	34	15	1	15	4	63
S15	16	31	9	1	20	1	26
$\%ee_{med}$	16	31	15	69	74	83	93

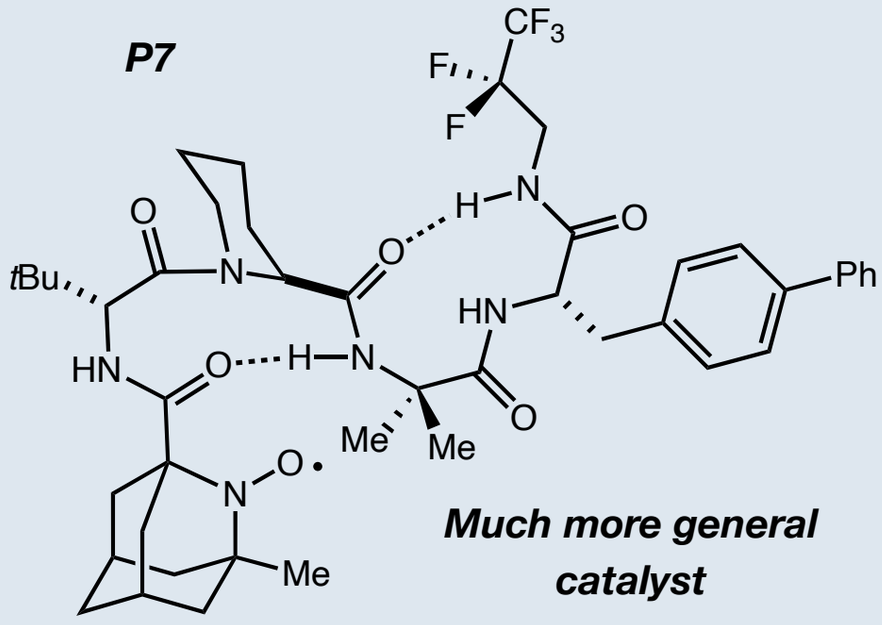
S1





P4

P7



Much more general catalyst

Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. *Science* **2023**, *380*, 701–712

Additional Reading

Reaction Generalization

Kim, H.; Gerosa, G.; Aronow, J.; Kasaplar, P.; Ouyang J.; Lingau, J .B.; Guerry, P.; Fares, C.; List, B. *Nat. Comm.* **2019**, *10.*

Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. *J. Am. Chem. Soc.* **2023**, *145*, 12870–12883

Rana, D.; Pfluger, P. M.; Holter, N. P.; Tan, G.; Glorius, F. *ACS Cent. Sci.* **2024**, *10*, 899–906

Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. *Nature* **2022**, *610*, 680–686

Miniaturization

Gorski, B.; Rein, J.; Norris, S.; Ji, Y.; McEuen, P. L.; Lin, S. *Nature*, **2025**, *637*, 354–361

Gesmundo, N.; Dykstra, K.; Douthwaite, J. L.; Kao, Y.; Zhao, R.; Mahjour, B.; Ferguson, R.; Dreher, S. Sauvagnat, B.; Sauri, J.; Cernak, T. *Nature Synthesis*, **2023**, *2*, 1082–1091

Esguevillas, M.; Fernandez, D. F.; Rincon, J. A.; Barberis, M.; Frutos, O. D.; Mateos, C.; Cerreda, S.; Agejas, J.; MacMillan, D. W. C. *J. Am. Chem. Soc.* **2021**, *7*, 1126–1134

Chemputer/Accelerated Serendipity

Robbins, D. W.; Hartwig, J. F. *Science*, **2011**, *333*, 1423–1427

Granda, J M.; Donina, L.; Dragone, V.; Long, D. Cronin, L. *Nature*, **2018**, *559*, 377–381

Collins, K. D.; Mensch, T.; Glorius, F. *Nat. Chem.* **2014**, *6*, 859–871

Modern approaches to methods development

Modern Paradigms in Screening

- *Reaction generalization*
- *“Accelerate” Serendipity*
- *Miniaturization of unique reaction set ups*

Data Science

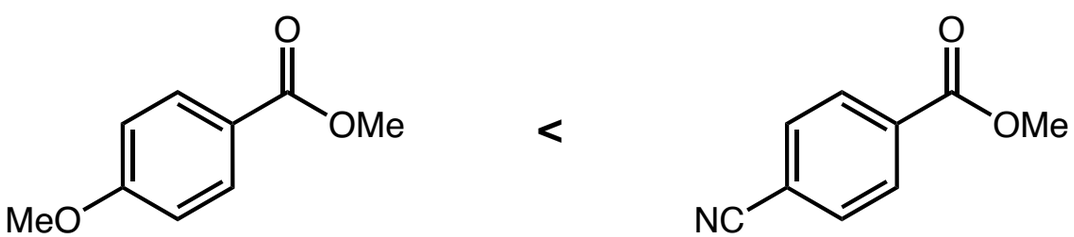
- *Catalyst optimization*
- *Predicting selectivity*
- *Discovery of new catalysts*

Machine Learning

- *What is machine learning*
- *Prediction of optimal conditions*
- *Selectivity prediction for complex systems*
- *Catalyst Discovery*

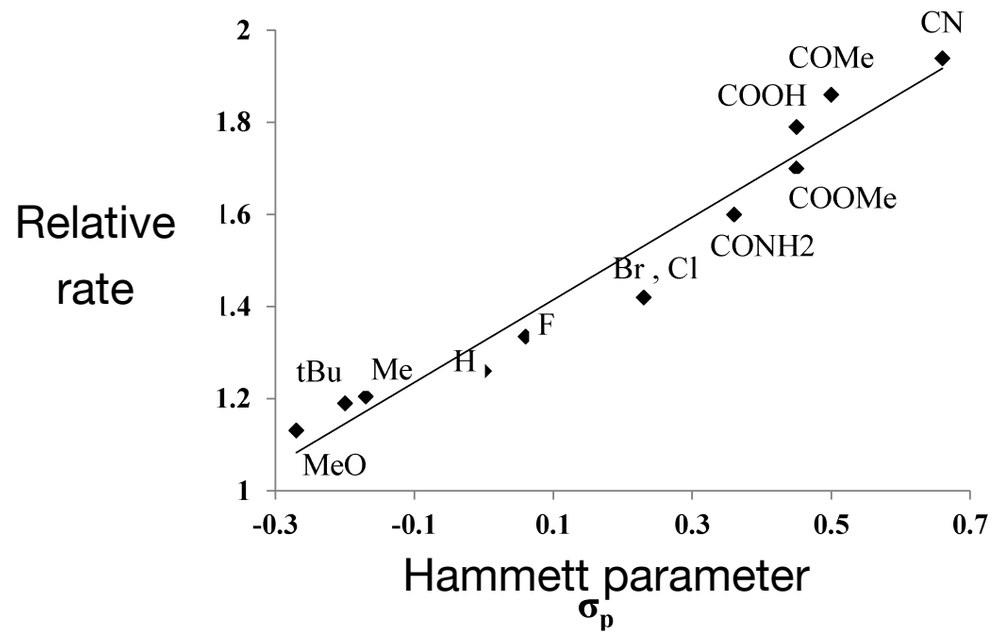
Linear Free Energy Relationships

Correlation of data to make mechanistic conclusions



Slower hydrolysis *Faster hydrolysis*

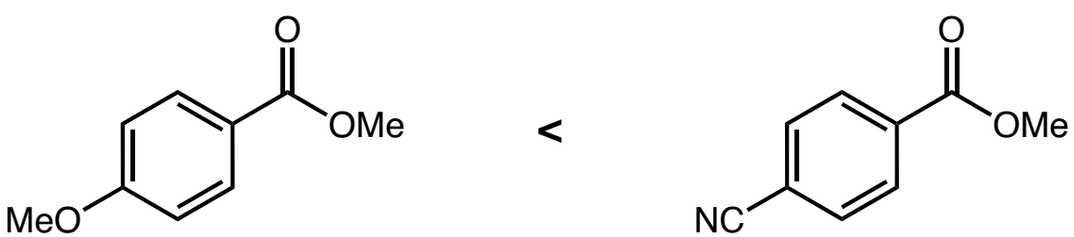
Hammett plot - 1935



Hammett, L. P.; *J. Am. Chem. Soc.* **1935**, *59*, 96–103

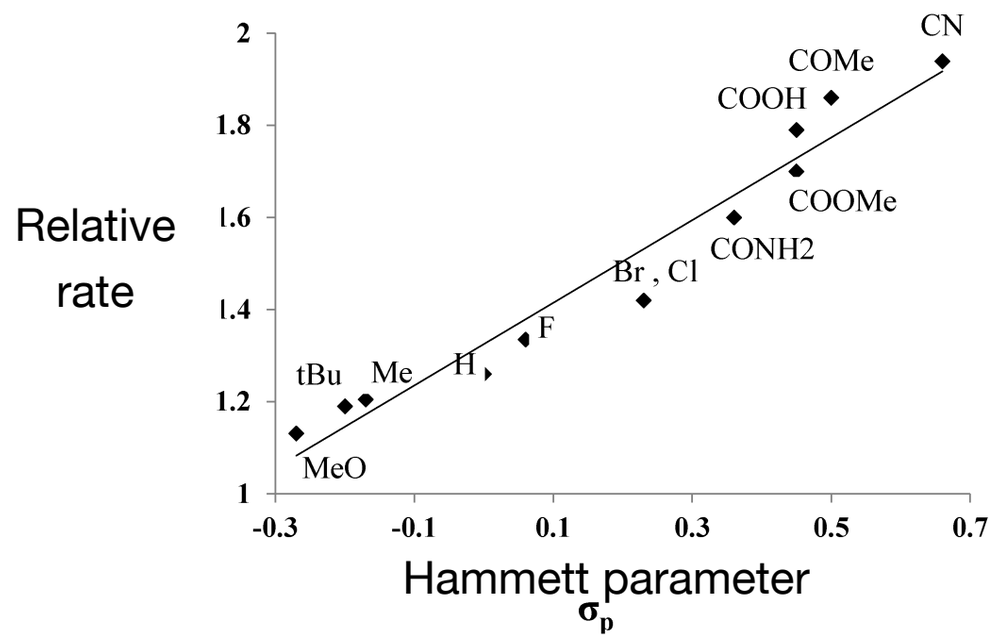
Linear Free Energy Relationships

Correlation of data to make mechanistic conclusions



Slower hydrolysis *Faster hydrolysis*

Hammett plot - 1935

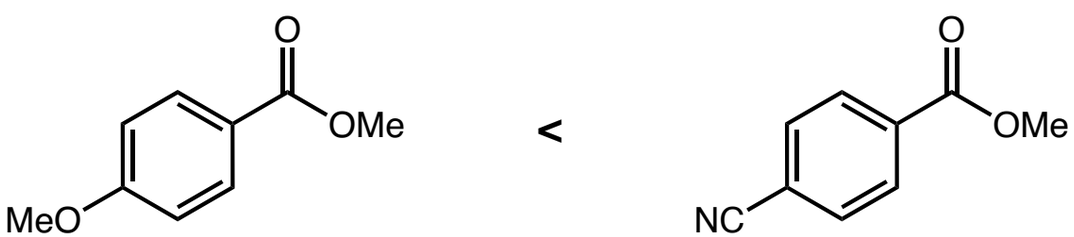


One of the most widely used “mechanistic probes” with **~66,600 publications mentioning Hammett plots**

Hammett, L. P.; *J. Am. Chem. Soc.* **1935**, *59*, 96–103

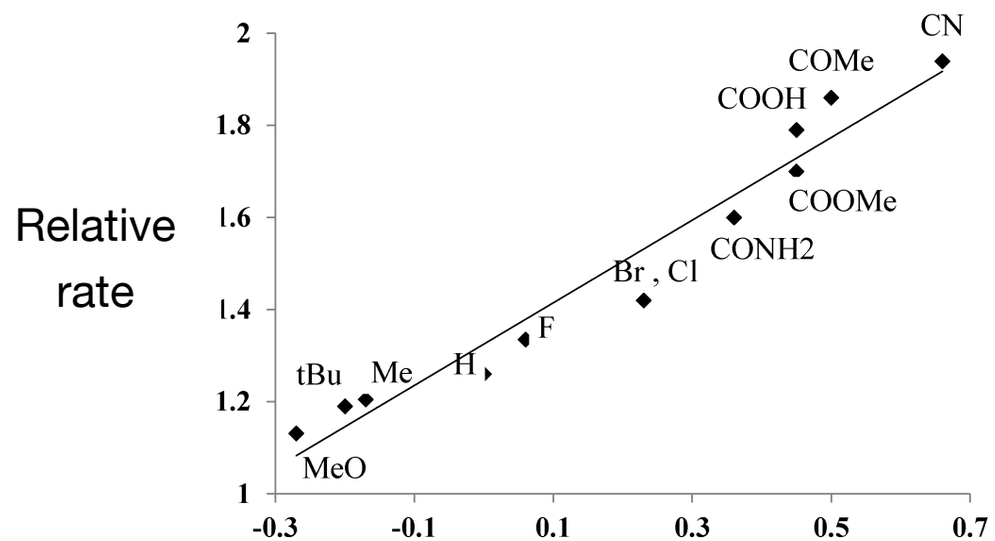
Linear Free Energy Relationships

Correlation of data to make mechanistic conclusions



Slower hydrolysis *Faster hydrolysis*

Hammett plot - 1935



Basic Hammett equation

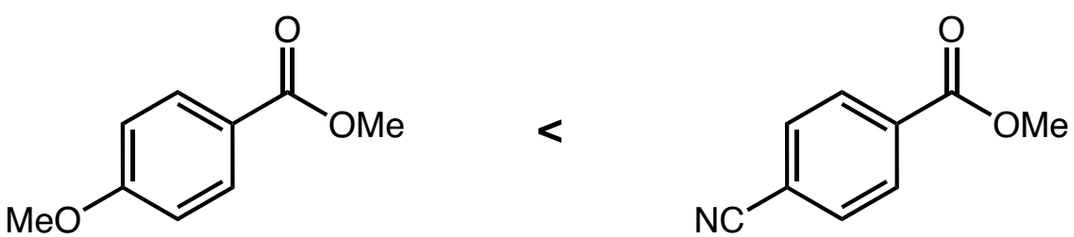
$$log(\frac{K}{K_o}) = \sigma\rho$$

Linear free-energy relation ship between relative rate and hammett parameter

Hammett, L. P.; *J. Am. Chem. Soc.* **1935**, *59*, 96–103

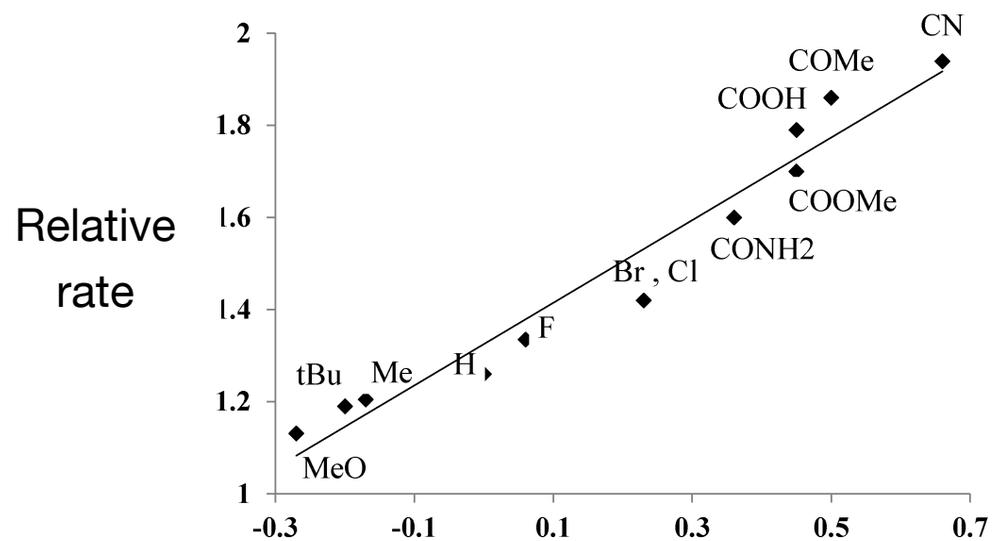
Linear Free Energy Relationships

Correlation of data to make mechanistic conclusions



Slower hydrolysis *Faster hydrolysis*

Hammett plot - 1935



<u>Basic Hammett equation</u>

$$log(\frac{K}{K_o}) = \sigma\rho$$

Linear free-energy relation ship between relative rate and hammett parameter

<u>Many other parameters</u>

Sterimol

Charton

Taft

A-values

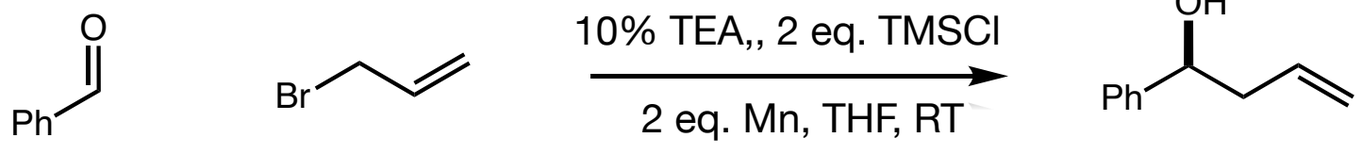
electric field

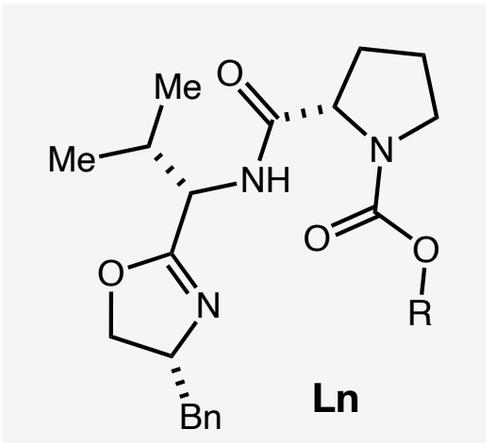
Ect.

Hammett, L. P.; *J. Am. Chem. Soc.* **1935**, *59*, 96–103

Linear Free Energy Relationships - Exceptions

Enantioselective Allylation

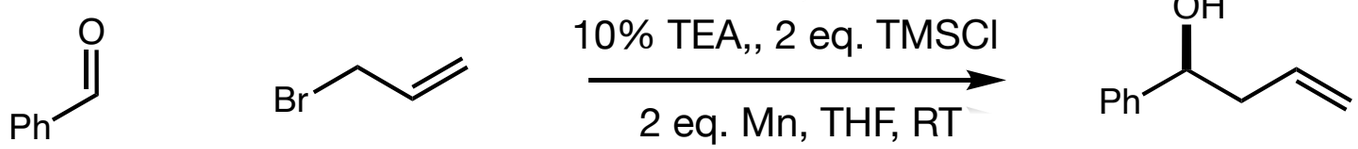


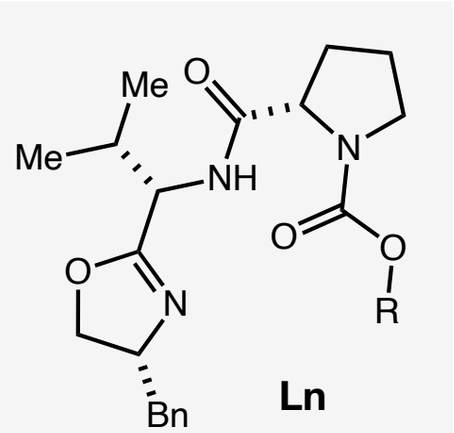


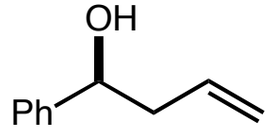
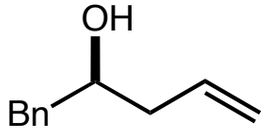
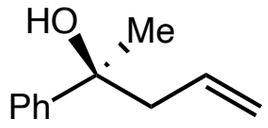
Miller, J. J.; Sigman, M. S.; *Angew. Chemie* **2008**, *47*, 771–774

Linear Free Energy Relationships - Exceptions

Enantioselective Allylation





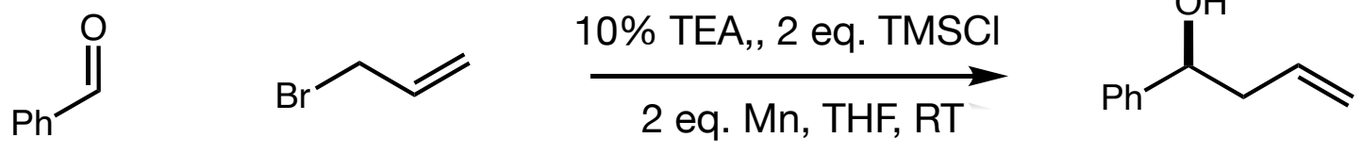
			
R = Me	1.5 e.r.	1 e.r.	0.3 e.r.
R = Et	1.9 e.r.	1.1 e.r.	0.3 e.r.
R = *i*-Pr	3.5 e.r.	1.3 e.r.	0.6 e.r.
R = *t*-Bu	23 e.r.	2.8 e.r.	2.2 e.r.
R = Ad	23 e.r.	2.8 e.r.	2.6 e.r.

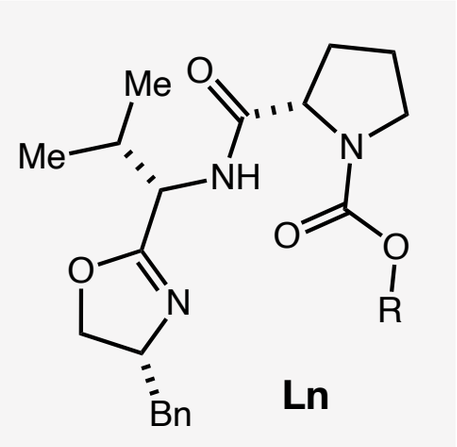
Linear free-energy relationship between log (e.r.) and Steric parameters

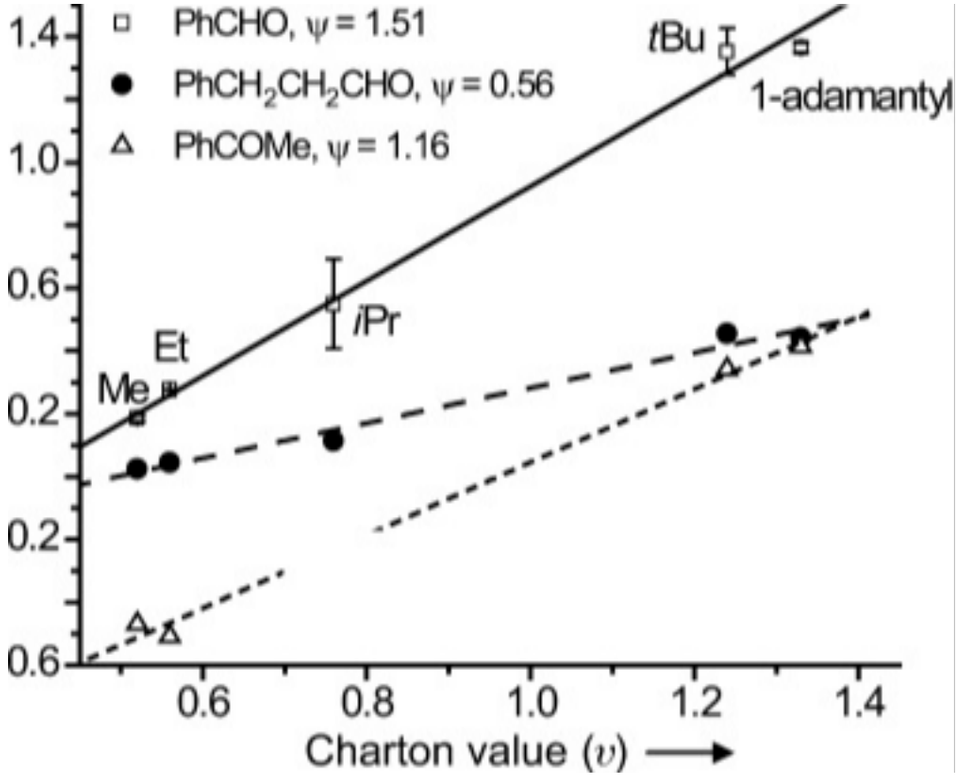
Miller, J. J.; Sigman, M. S.; *Angew. Chemie* **2008**, *47*, 771–774

Linear Free Energy Relationships - Exceptions

Enantioselective Allylation





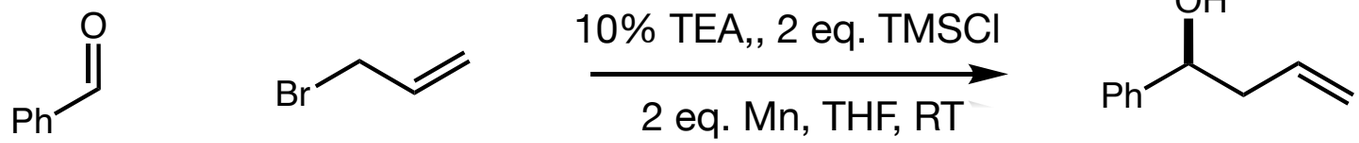


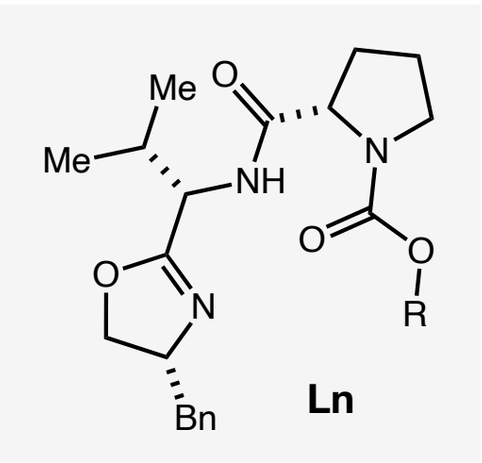
Linear free-energy relationship between log (e.r.) and Steric parameters

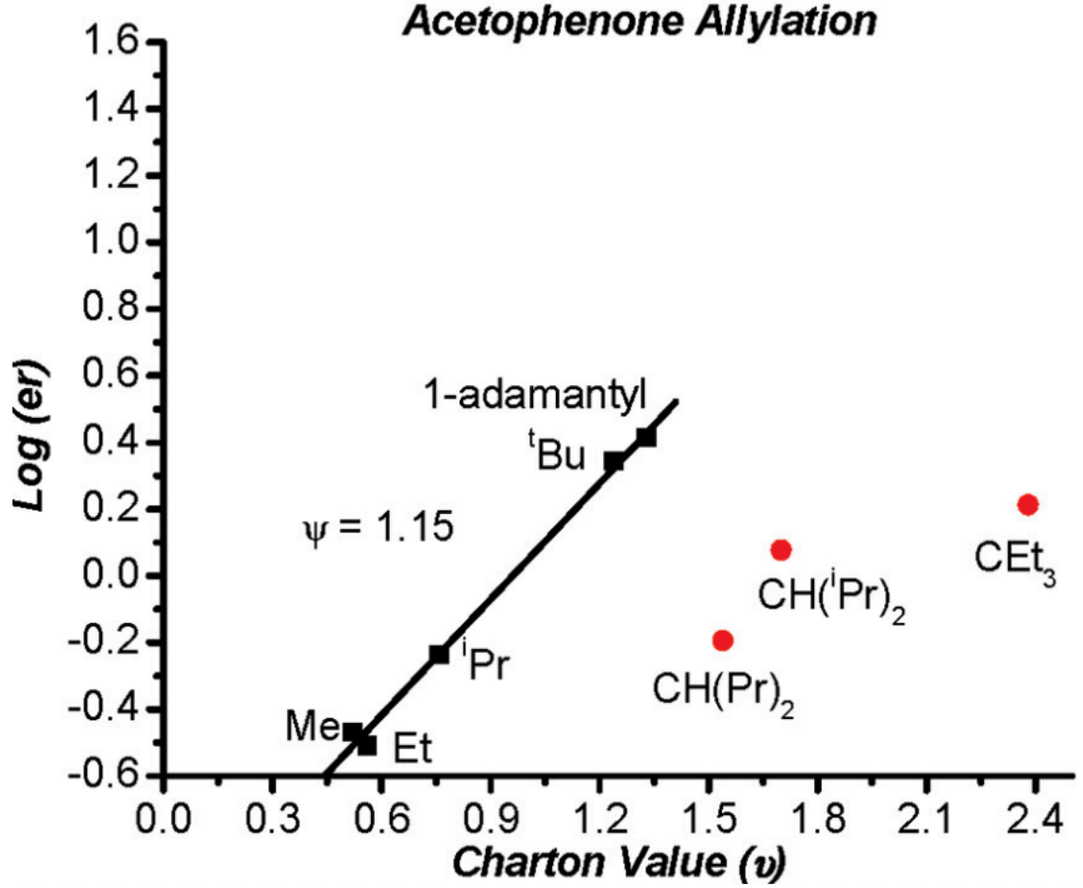
Miller, J. J.; Sigman, M. S.; *Angew. Chemie* **2008**, *47*, 771–774

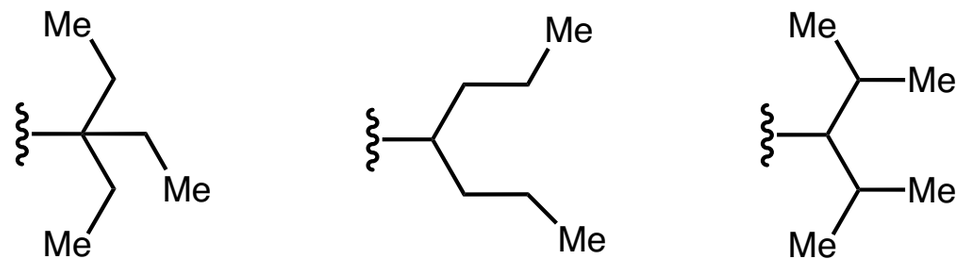
Linear Free Energy Relationships - Exceptions

Enantioselective Allylation





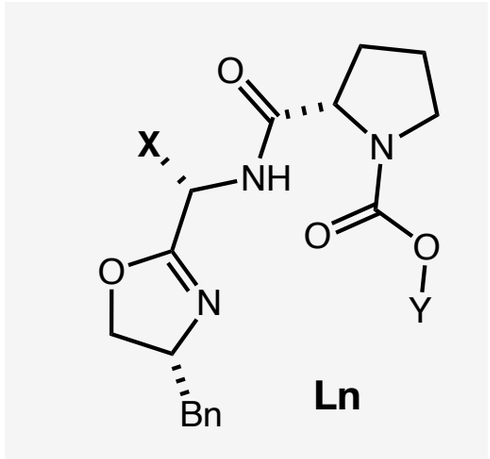




Even bigger Charton Values give worse e.r., breaks the linear relationship

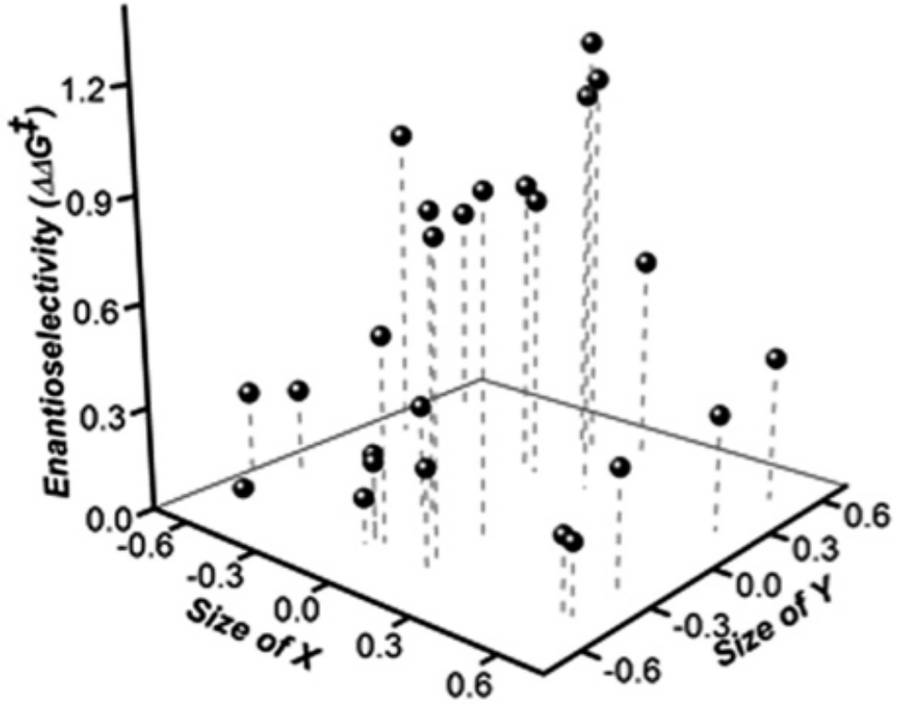
Miller, J. J.; Sigman, M. S.; *J. Org. Chem.* **2009**, *74*, 7633–7643

Multivariate Linear Free Energy Relationships



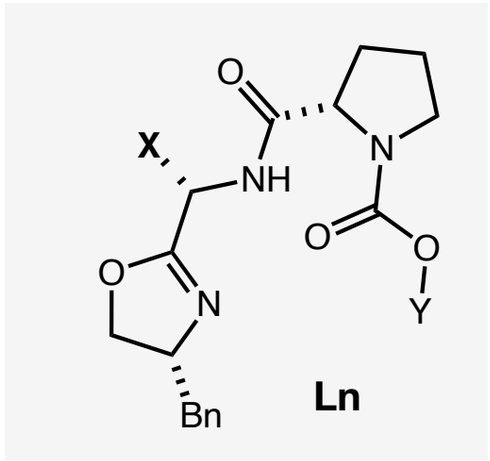
X =	**Y =**
H	*H*
Me	*Me*
Et	*iPr*
iPr	*tBu*
tBu	*CH(Pr)$_2$*

25 ligand library to determine relation ship between steric bulk at X and Y



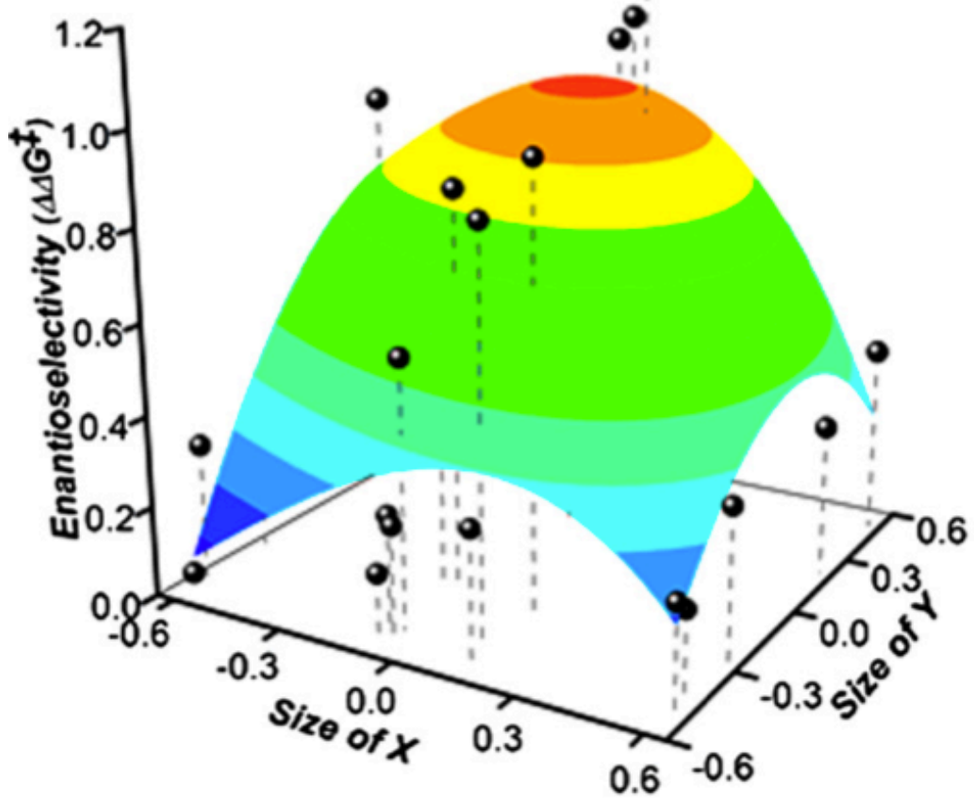
Harper, K. C.; Sigman, M. S.; *PNAS* **2011**, *108*, 2179–2183

Multivariate Linear Free Energy Relationships



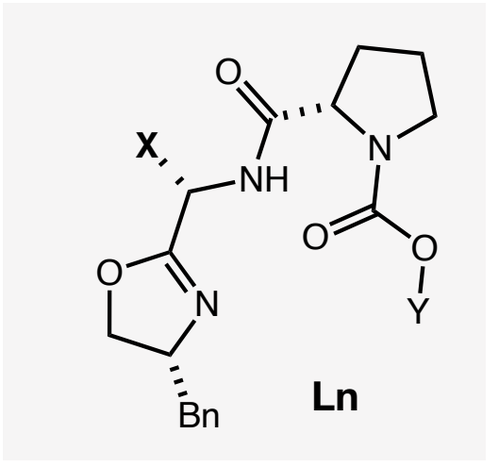
X =	**Y =**
H	*H*
Me	*Me*
Et	*iPr*
iPr	*tBu*
tBu	*CH(Pr)$_2$*

25 ligand library to determine relation ship between steric bulk at X and Y



Harper, K. C.; Sigman, M. S.; *PNAS* **2011**, *108*, 2179–2183

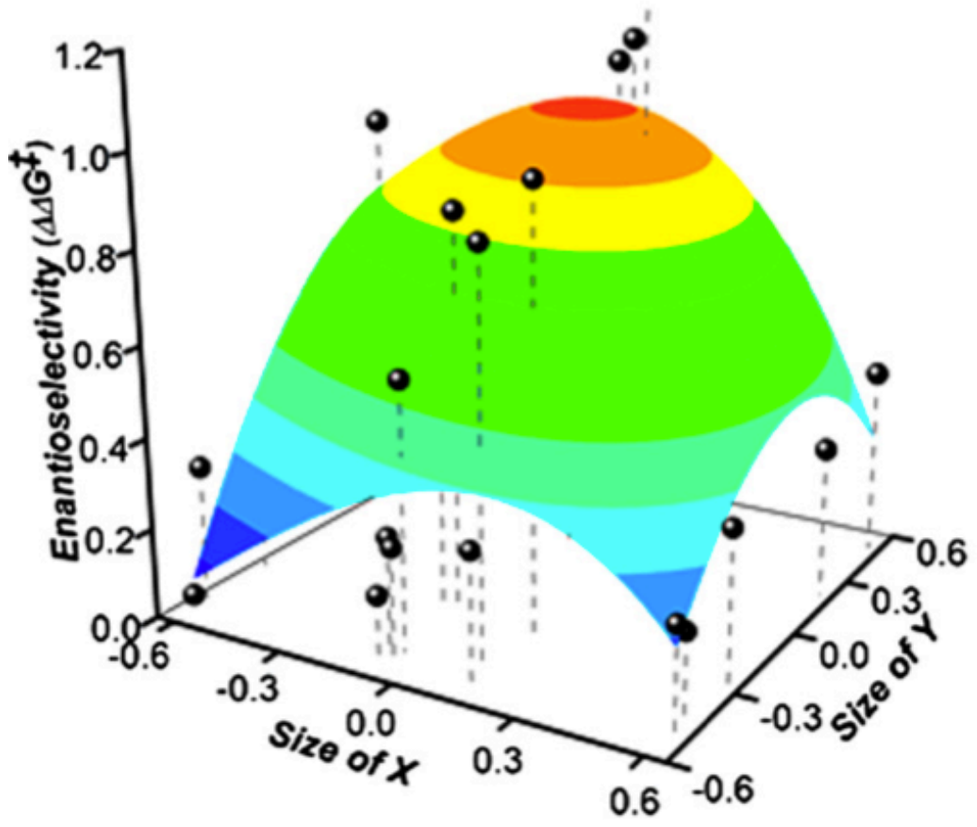
Multivariate Linear Free Energy Relationships

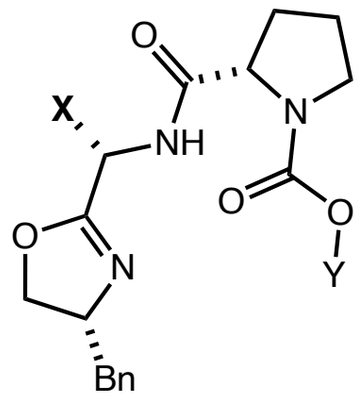


Ln

X =	**Y =**
H	*H*
Me	*Me*
Et	*iPr*
iPr	*tBu*
tBu	*CH(Pr)$_2$*

25 ligand library to determine relation ship between steric bulk at X and Y





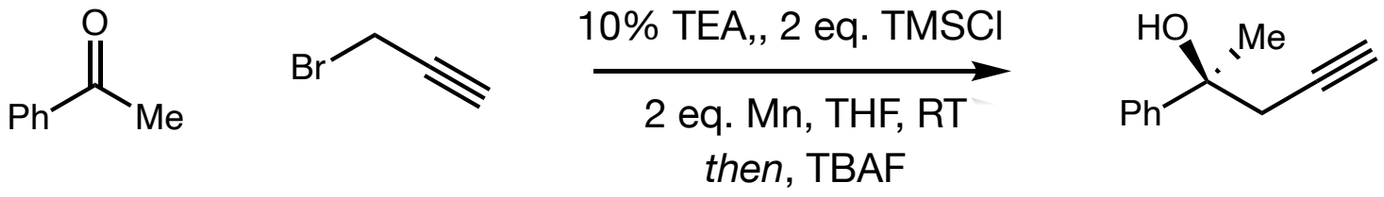
Relatively accurate predictor even for much bigger Y groups

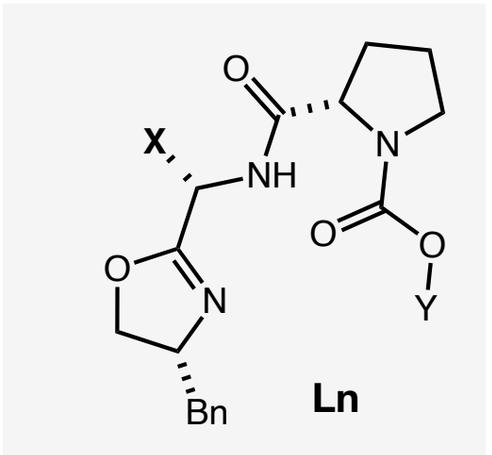
X =	**Y =**	**Predicted e.r.**	**Measured e.r.**
H	*Me*	*49:51*	*49:51*
Me	*tBu*	*44:56*	*46:54*
t-Bu	*C(Et)$_3$*	*40:60*	*42:58*
iPr	*C(Et)$_3$*	*49:51*	*54:46*

Harper, K. C.; Sigman, M. S.; *PNAS* **2011**, *108*, 2179–2183

Multivariate Linear Free Energy Relationships

Enantioselective Propargylation

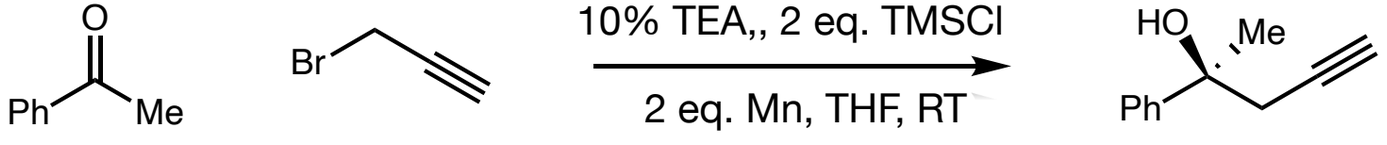


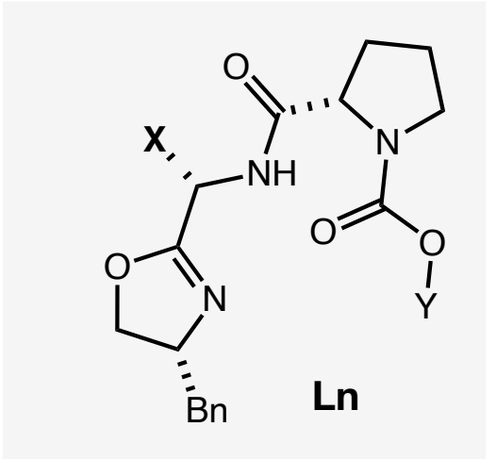


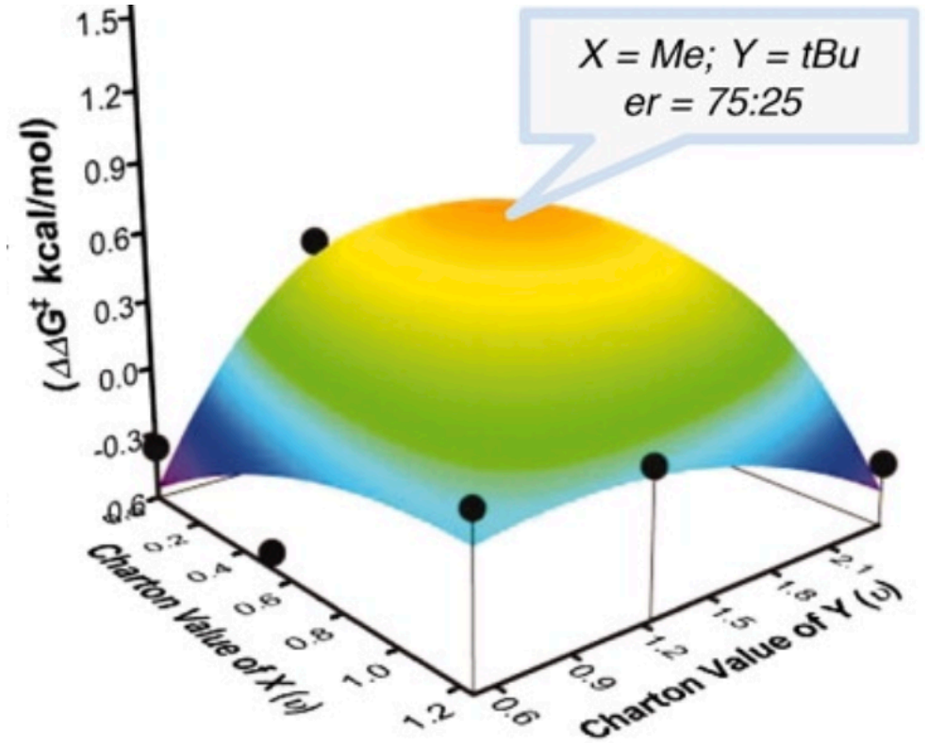
Harper, K. C.; Sigman, M. S.; *Science* **2011**, *333*, 1875–1878

Multivariate Linear Free Energy Relationships

Enantioselective Propargylation





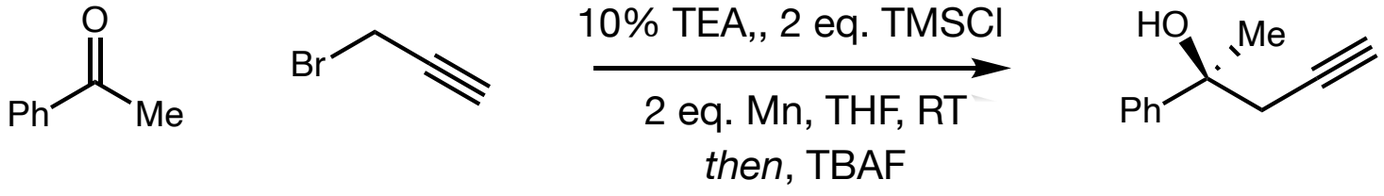


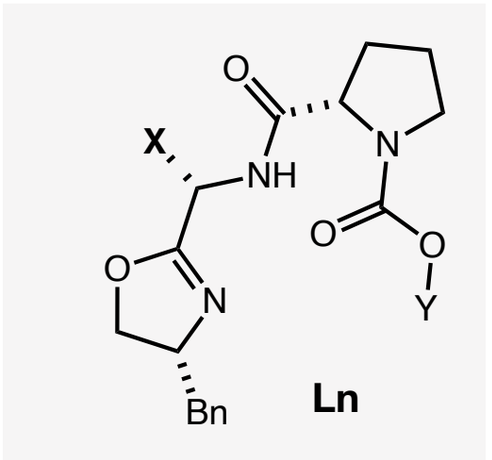
Low e.r. predicted and overall small correlation between e.r. and steric parameters

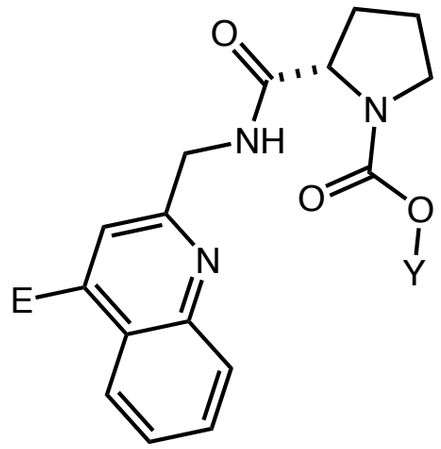
Harper, K. C.; Sigman, M. S.; *Science* **2011**, *333*, 1875–1878

Multivariate Linear Free Energy Relationships

Enantioselective Propargylation







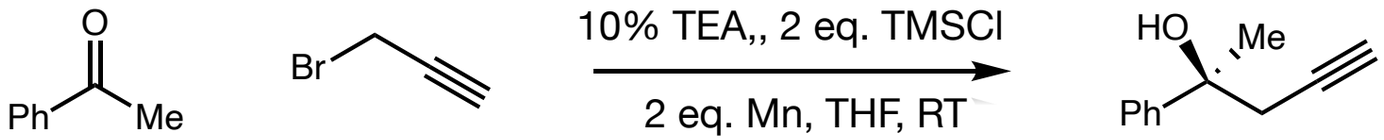
E = *Hammett parameter*

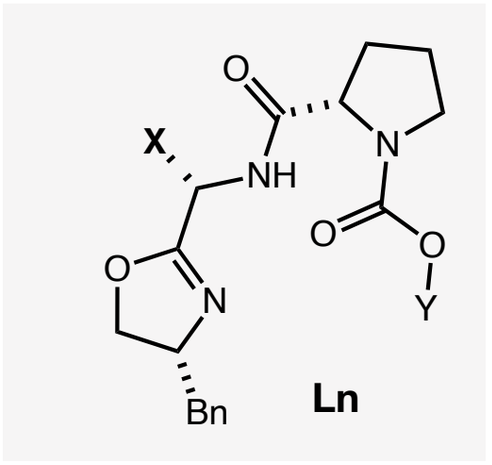
Y = *Steric parameter*

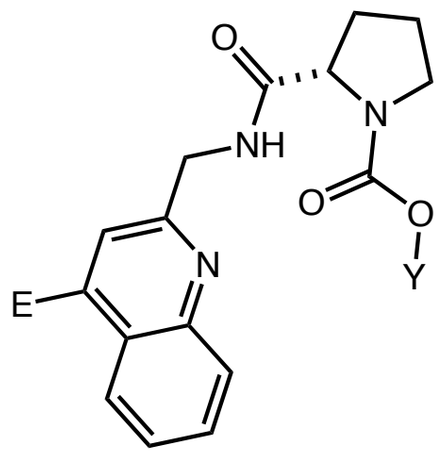
Harper, K. C.; Sigman, M. S.; *Science* **2011**, *333*, 1875–1878

Multivariate Linear Free Energy Relationships

Enantioselective Propargylation

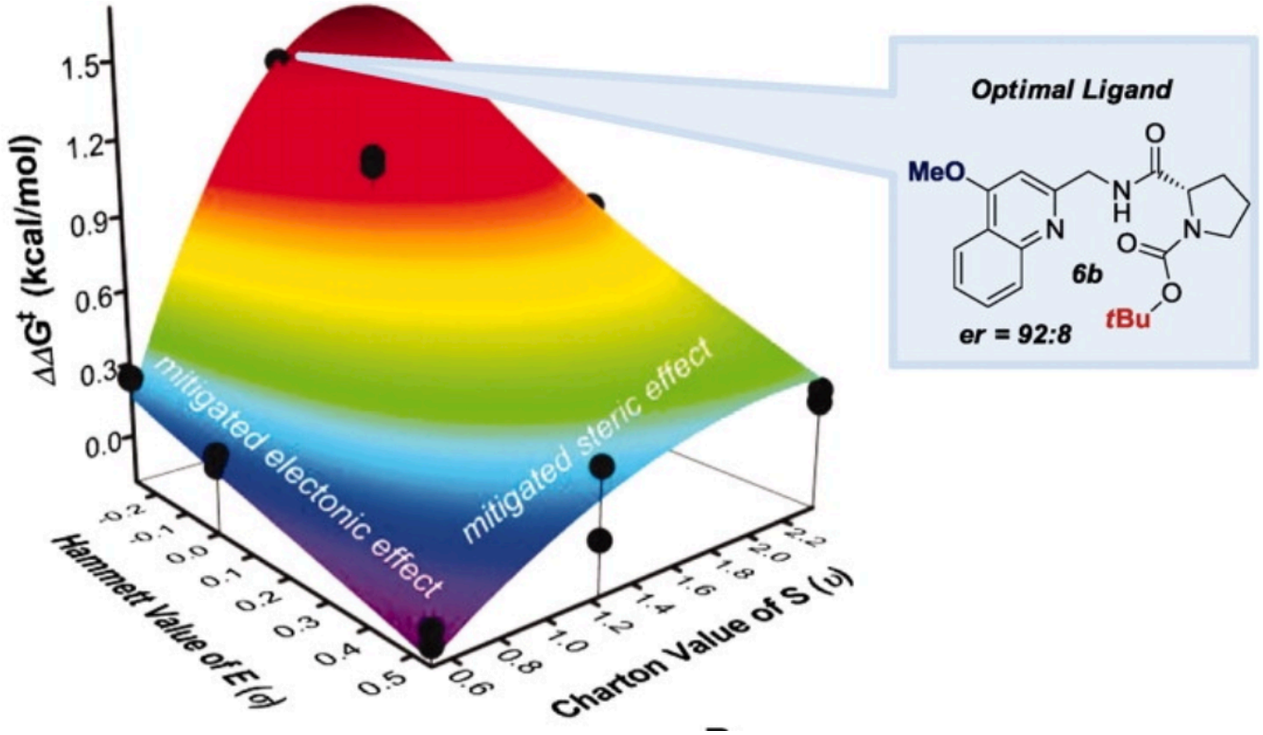






E = *Hammett parameter*

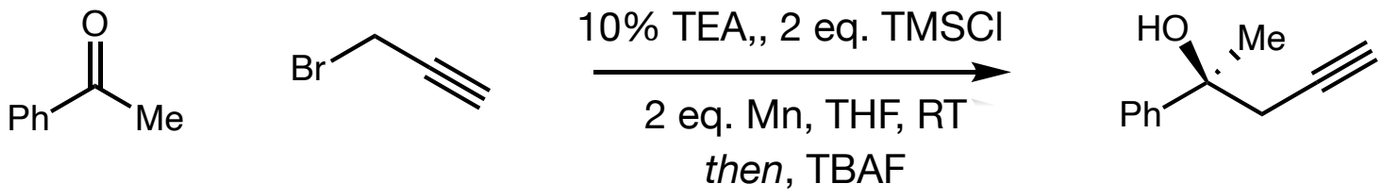
Y = *Steric parameter*

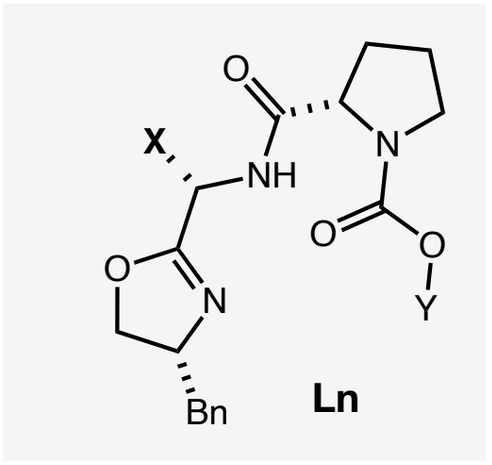


Harper, K. C.; Sigman, M. S.; *Science* **2011**, *333*, 1875–1878

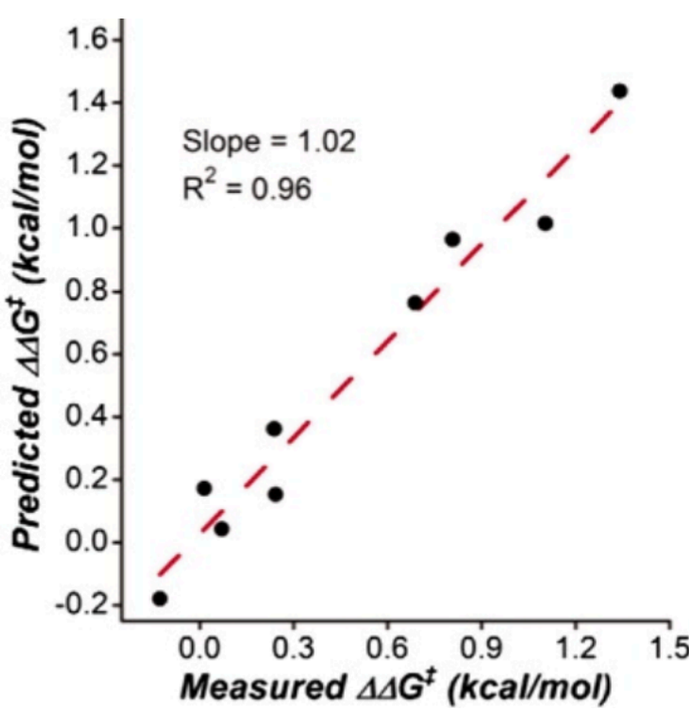
Multivariate Linear Free Energy Relationships

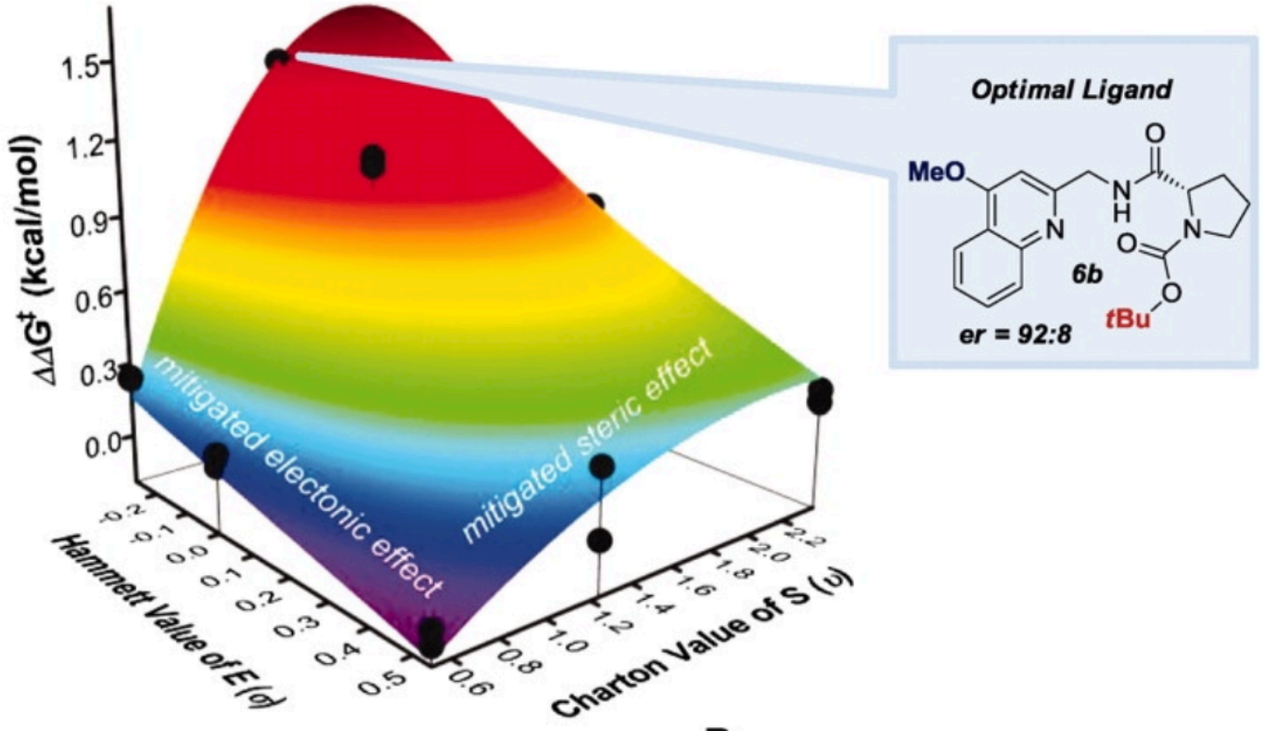
Enantioselective Propargylation





Model gives accurate predictions

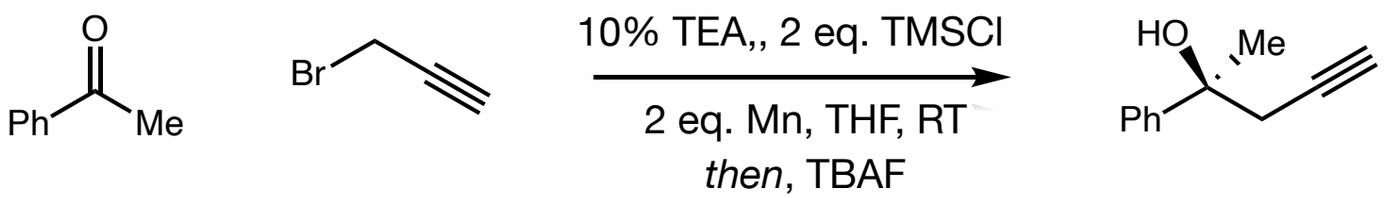


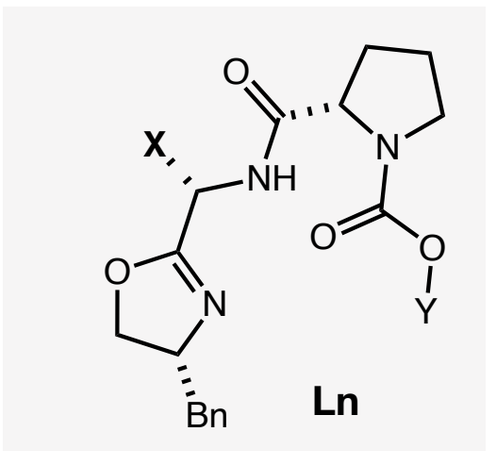


Harper, K. C.; Sigman, M. S.; *Science* **2011**, *333*, 1875–1878

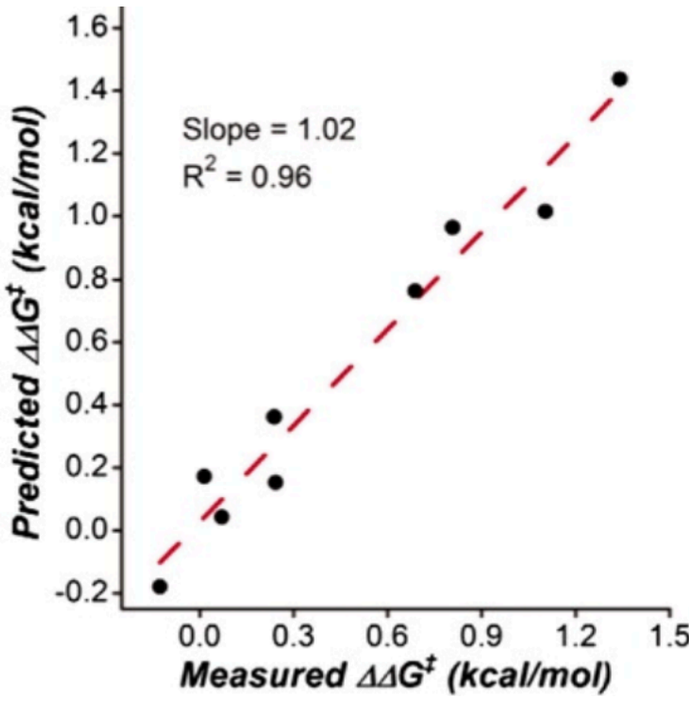
Multivariate Linear Free Energy Relationships

Enantioselective Propargylation





Model gives accurate predictions



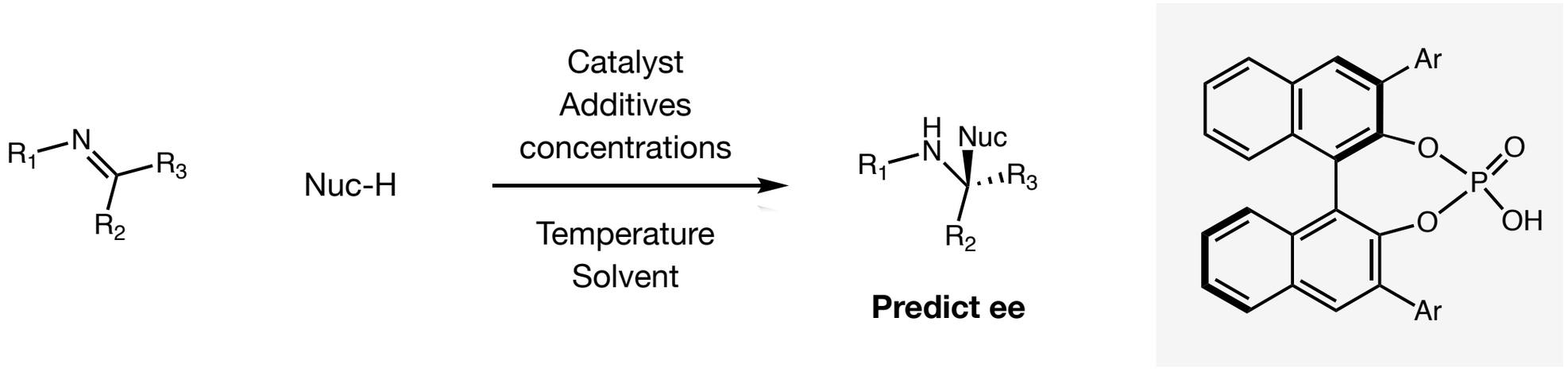
Mathematical formula

$$\Delta\Delta G^{\ddagger} = -1.20 + 1.22E + 2.84S - 0.85S^2 - 3.79\mathrm{ES} + 1.25\,\mathrm{ES}^2$$

Modeling can be done on more than just 2 parameters

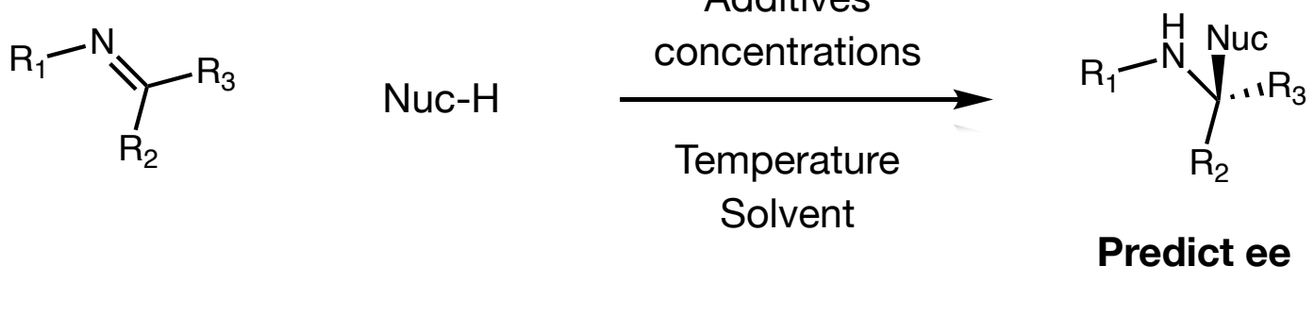
Harper, K. C.; Sigman, M. S.; *Science* **2011**, *333*, 1875–1878

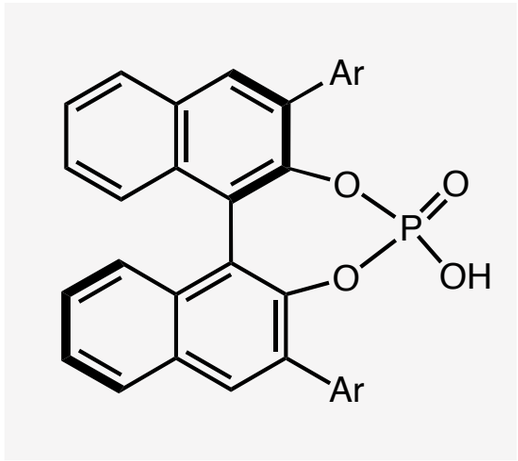
Multivariate Linear Free Energy Relationships - Predicting ee



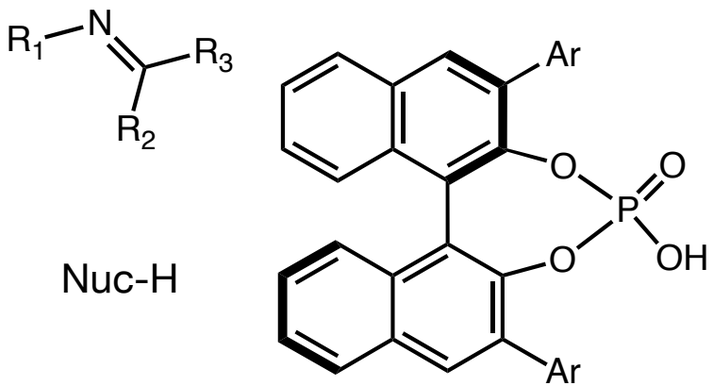
Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

Multivariate Linear Free Energy Relationships - Predicting ee





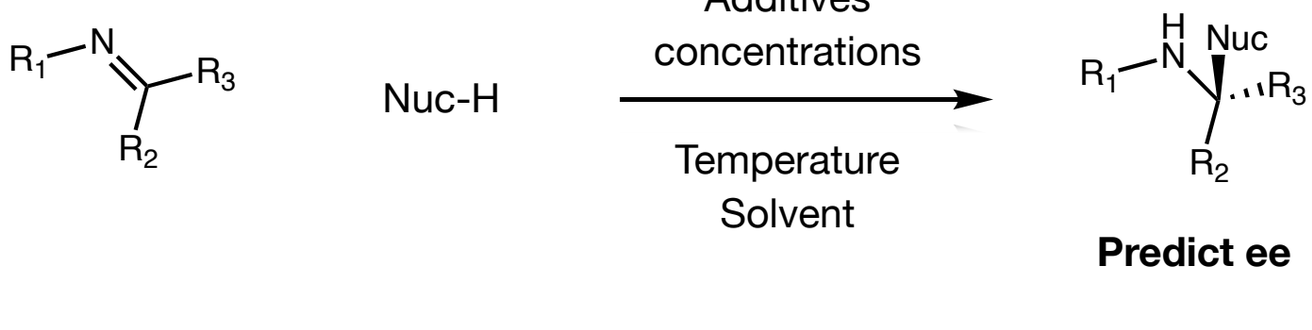
Parameterize all components



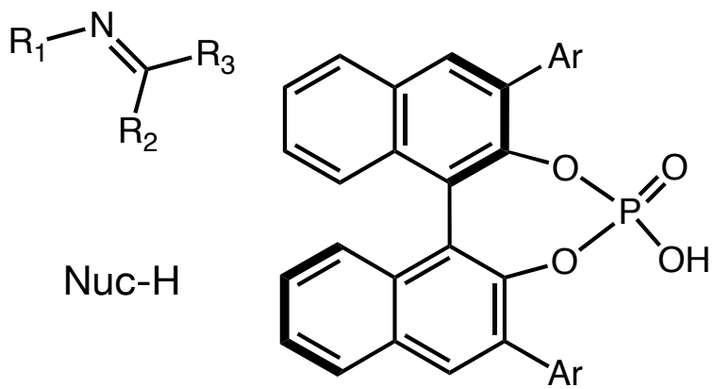
catalyst, additives, concentrations,
temperature, solvent

Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

Multivariate Linear Free Energy Relationships - Predicting ee



Parameterize all components



catalyst, additives, concentrations, temperature, solvent

Gather data on a training set

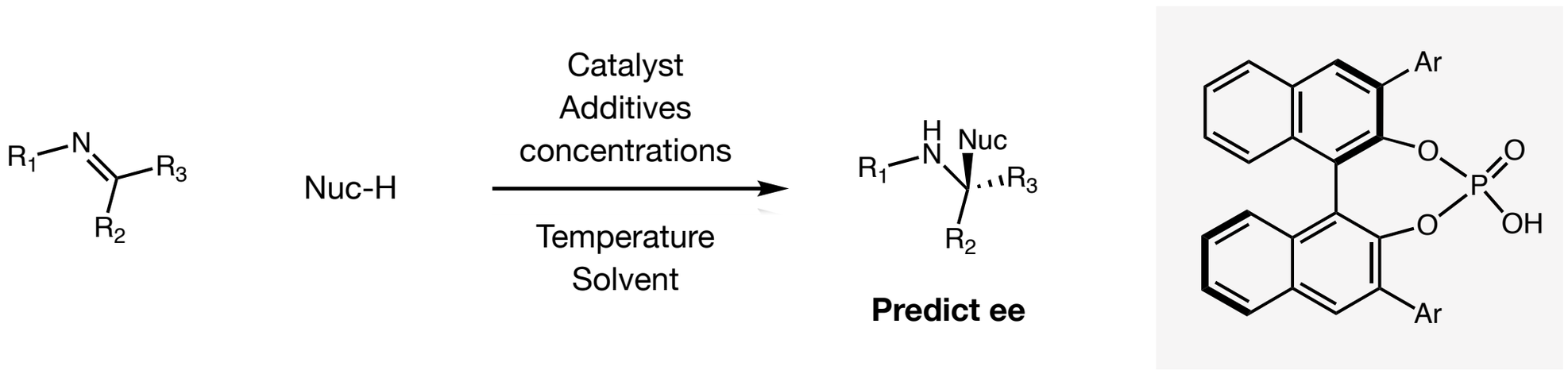
Analyze and model

1. Determine the parameters that have the highest effect on ee

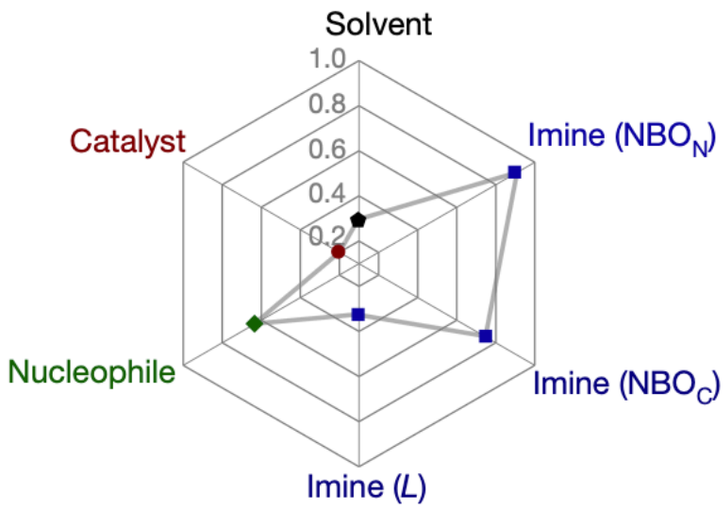
2. Derive a mathematical formula that predicts ee based on the important parameters

Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

Multivariate Linear Free Energy Relationships - Predicting ee

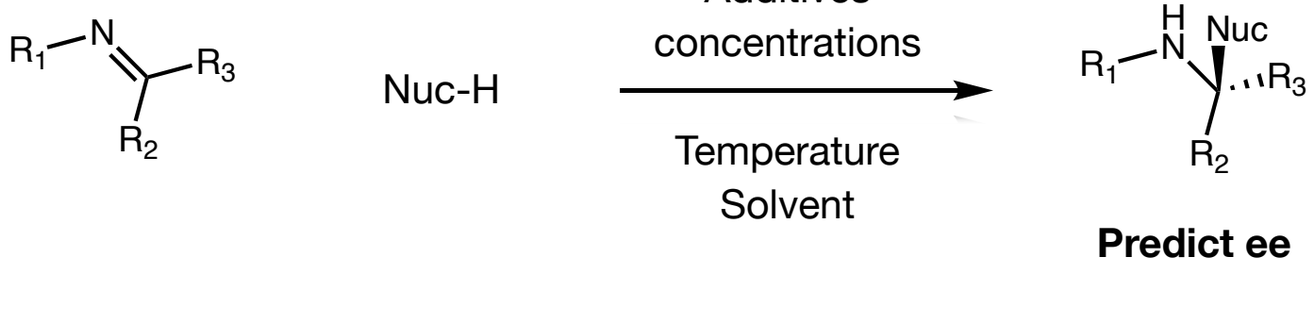


Determine and weigh important parameters

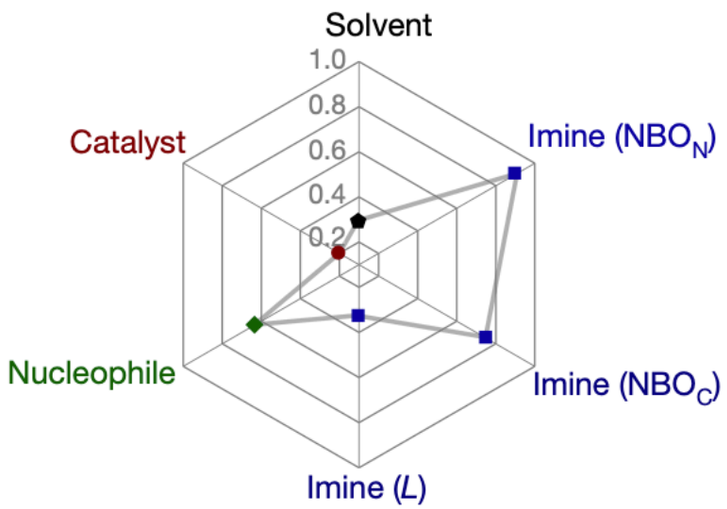


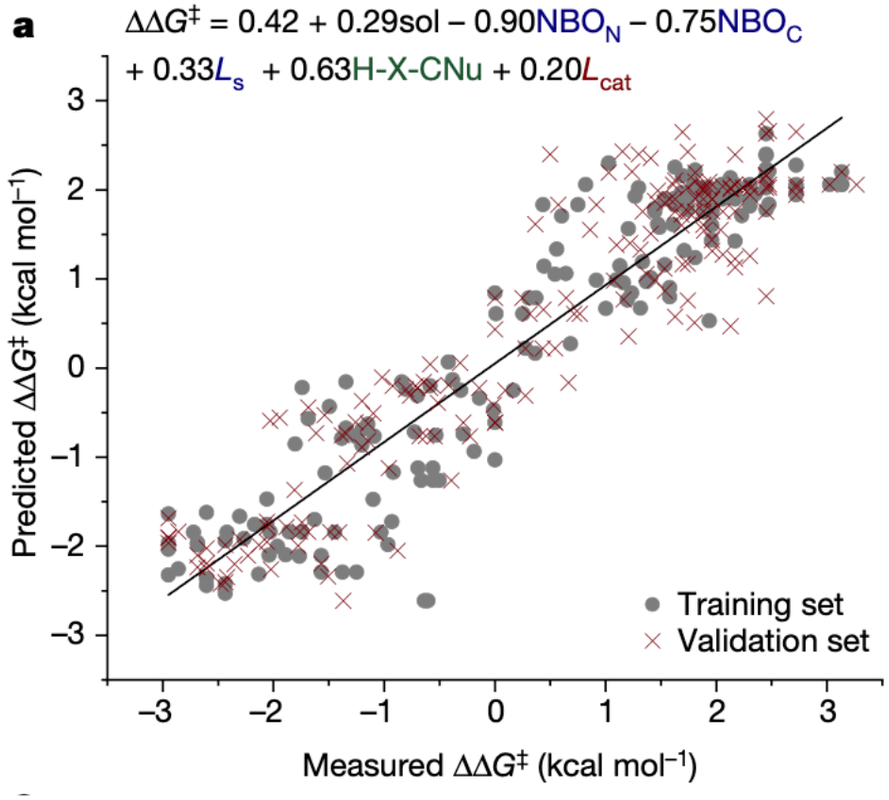
Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

Multivariate Linear Free Energy Relationships - Predicting ee



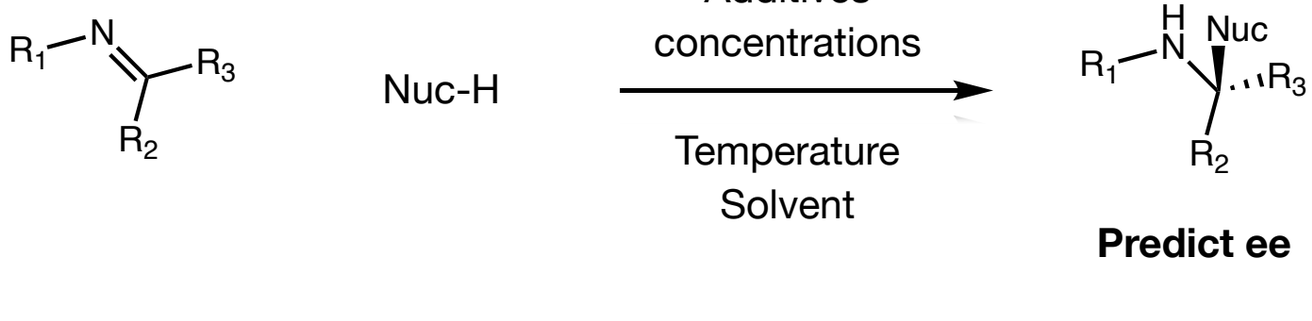
Determine and weigh important parameters

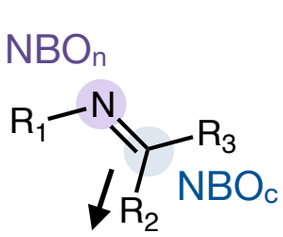




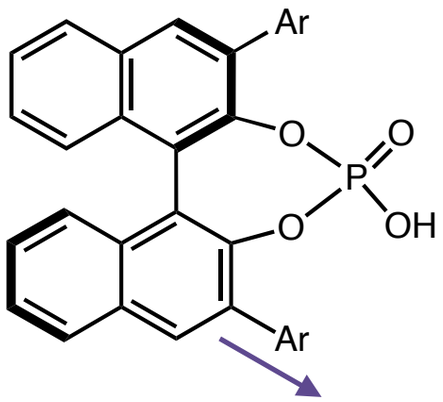
Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

Multivariate Linear Free Energy Relationships - Predicting ee



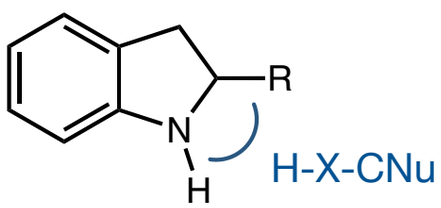


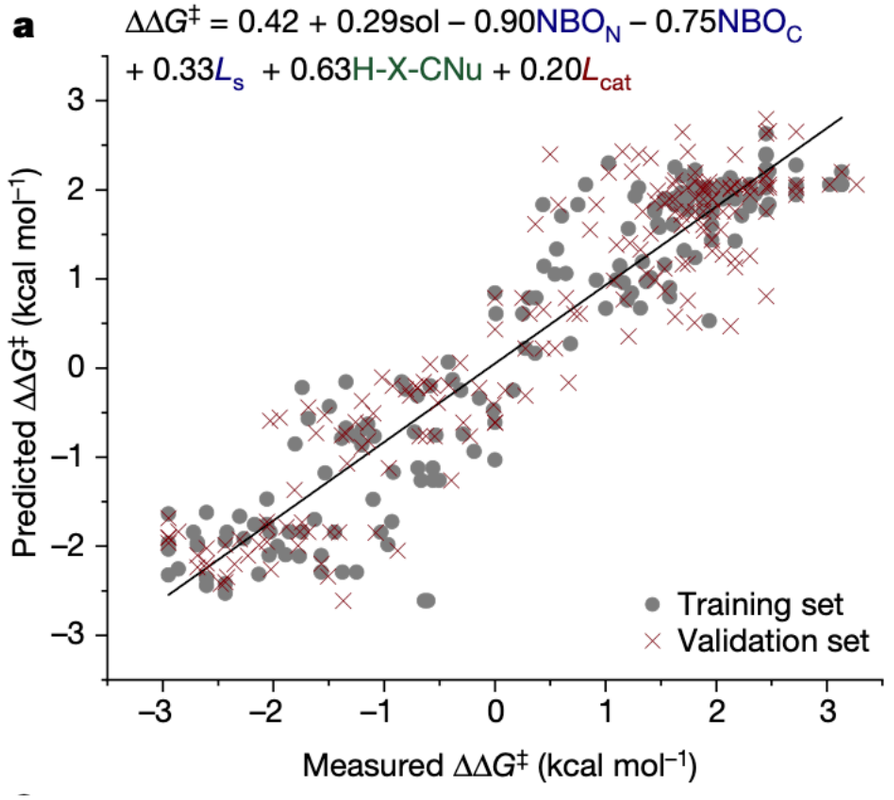
L_s: Length of smallest substituent



L_{cat}: Length of catalyst

sol = solvent term

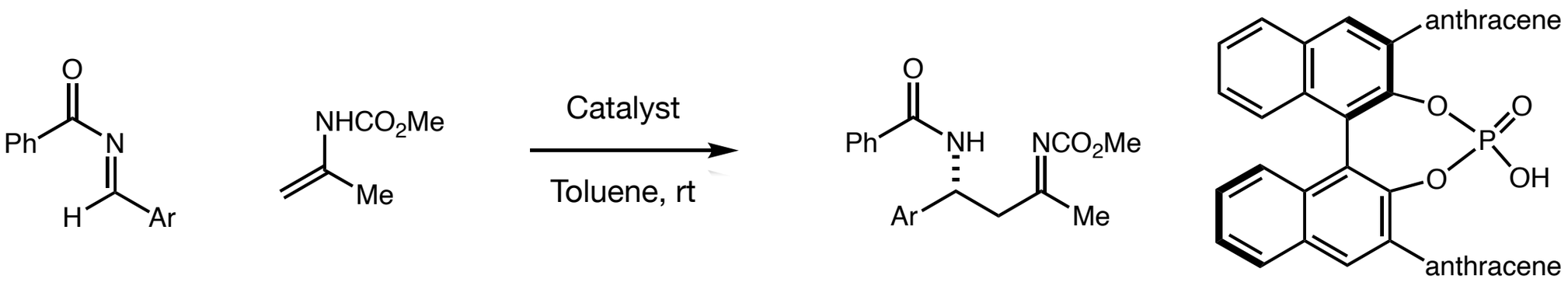




Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

Multivariate Linear Free Energy Relationships - Predicting ee

Testing the accuracy of the model with out-of-sample test substrates



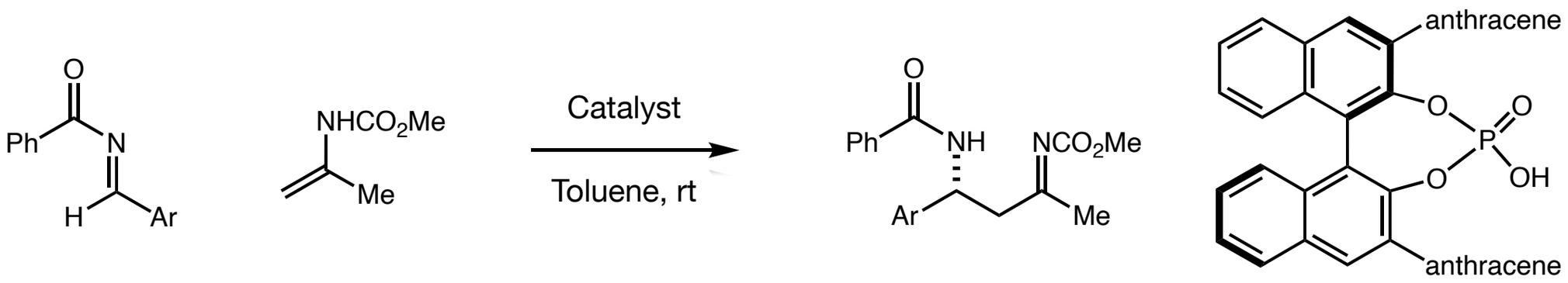
15 examples

All predicted within 5% ee

Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

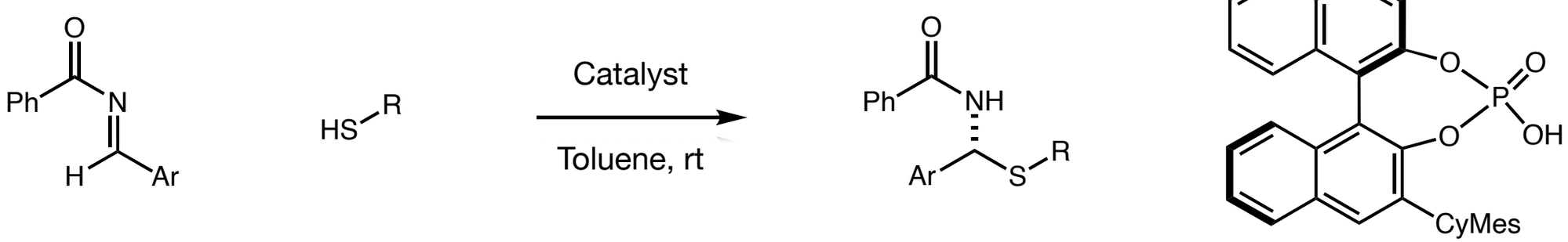
Multivariate Linear Free Energy Relationships - Predicting ee

Testing the accuracy of the model with out-of-sample test substrates



15 examples

All predicted within 5% ee

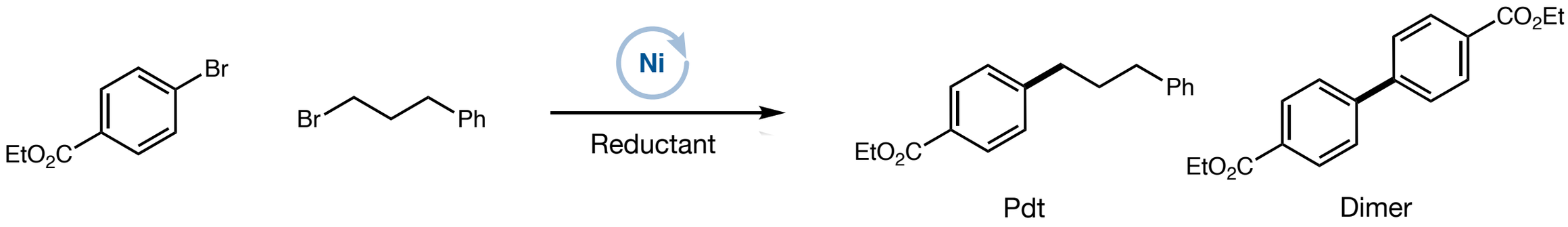


34 examples

26 examples within 5% ee

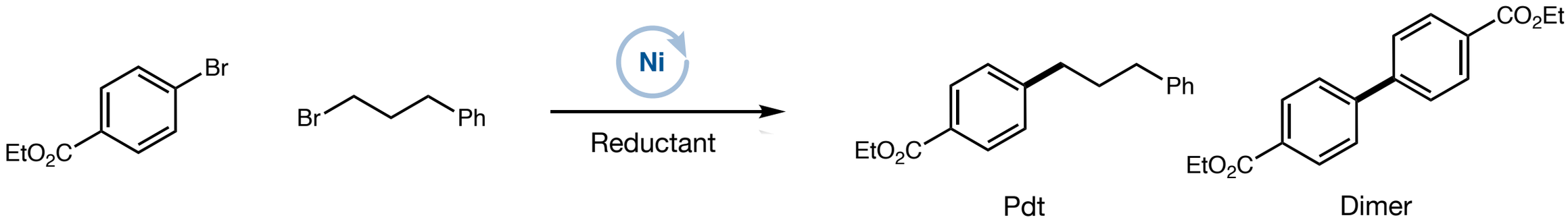
Reid, J. P.; Sigman, M. S.; *Nature* **2019**, *571*, 343 – 348

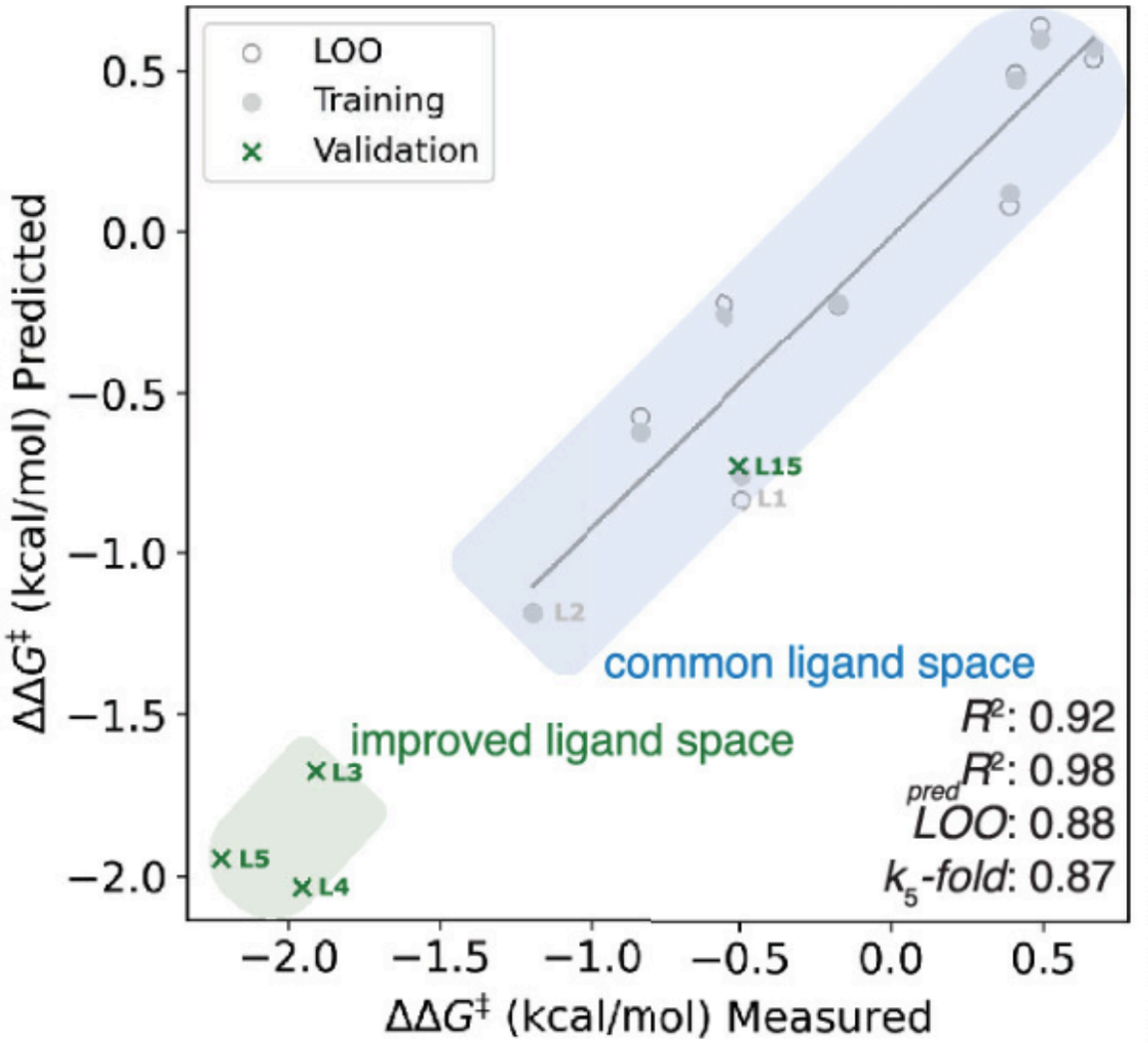
Multivariate Linear Free Energy Relationships - Finding New Catalysts



Akana, M. E.; Tcyrulnikov, S.; Scheider, B. D.; Reyes, G. P.; Monfette, S.; Sigman, M. S. Hansen, E. C.; Weix, D. J. *J. Am. Chem. Soc.* **2024**, *146*, 3043–3051

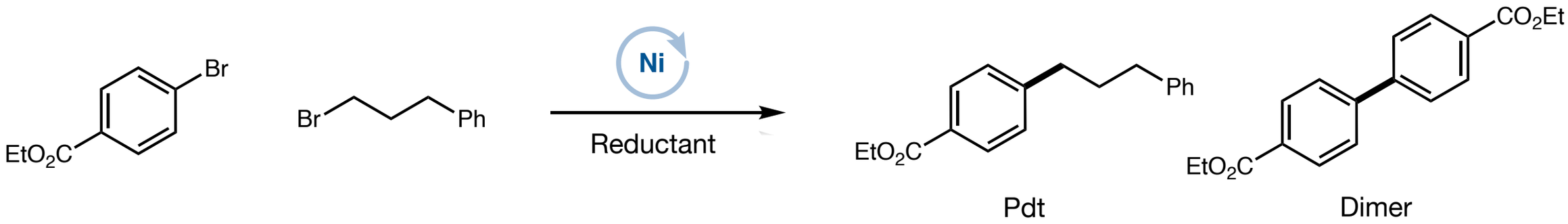
Multivariate Linear Free Energy Relationships - Finding New Catalysts





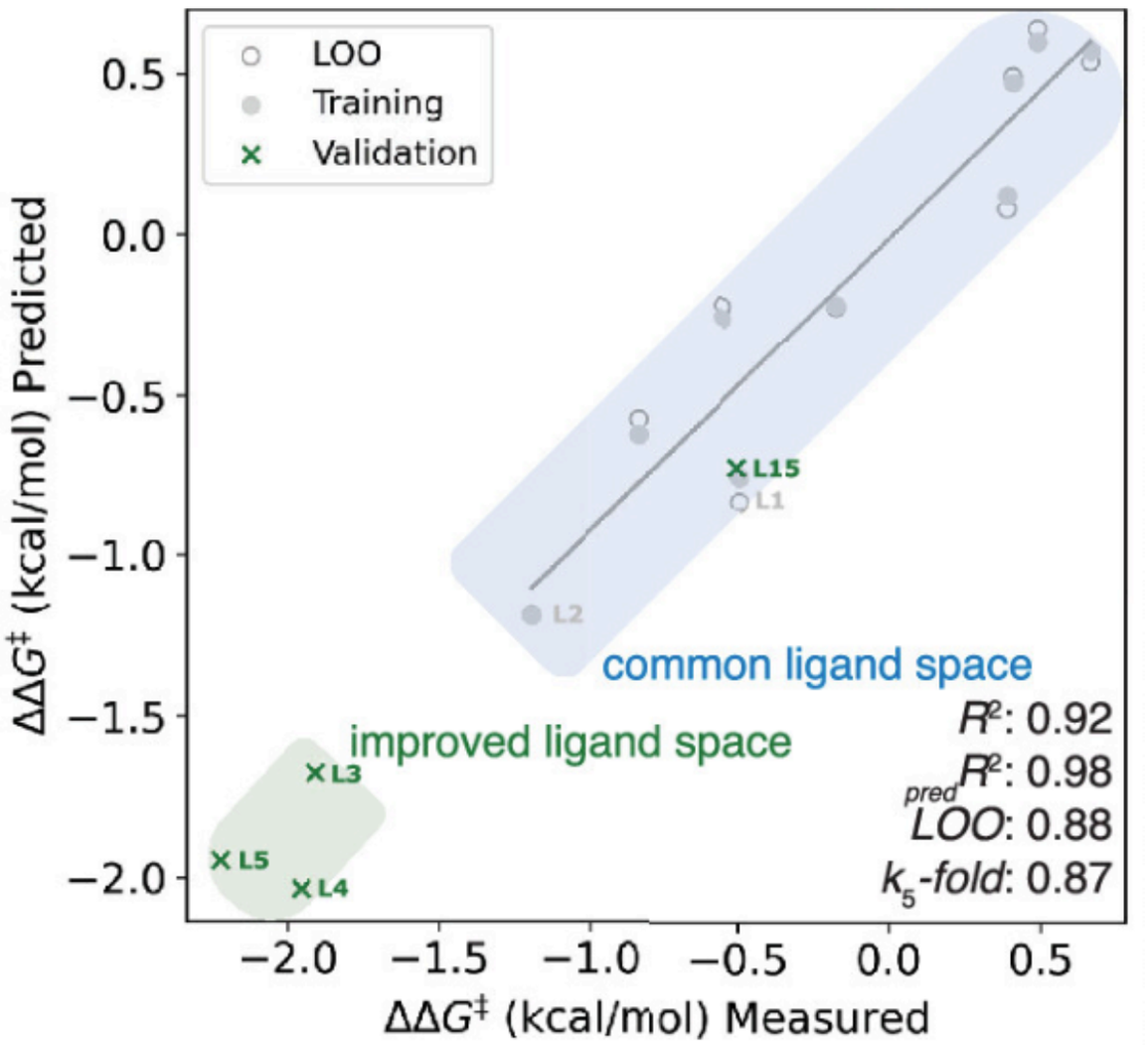
Akana, M. E.; Tcyrulnikov, S.; Scheider, B. D.; Reyes, G. P.; Monfette, S.; Sigman, M. S. Hansen, E. C.; Weix, D. J. *J. Am. Chem. Soc.* **2024**, *146*, 3043–3051

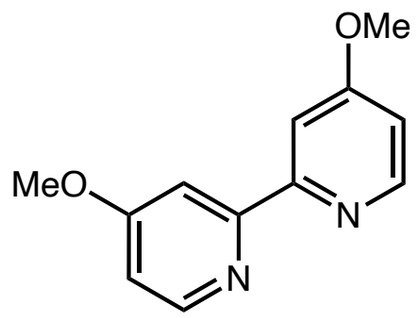
Multivariate Linear Free Energy Relationships - Finding New Catalysts



Pdt

Dimer

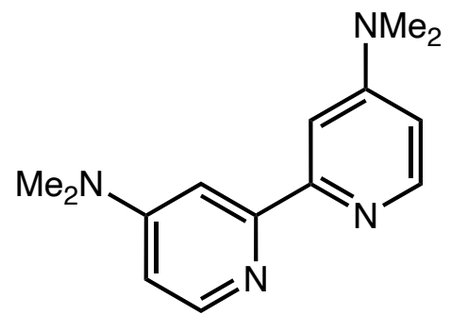




Previous best

61% yield

7:1 Pdt:dimer



Predicted best

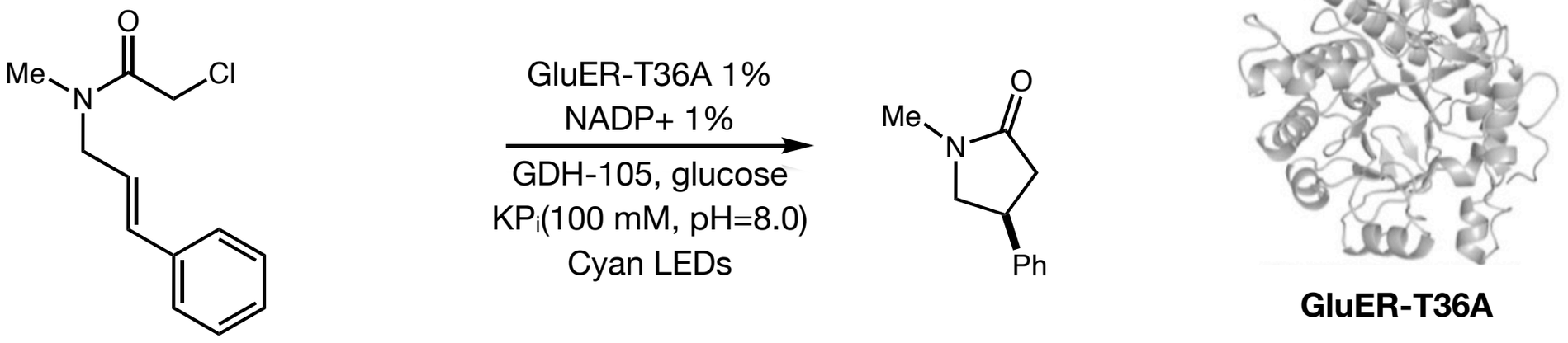
92% yield

187:1 Pdt:dimer

Poor understanding of the origin of the dimer, but model provides a hypothesis free solution to the problem

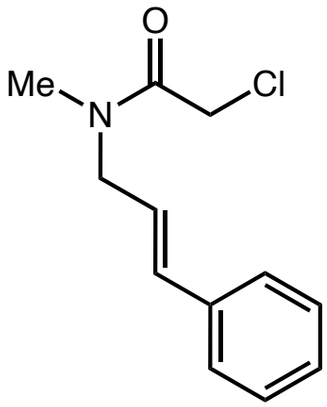
Akana, M. E.; Tcyrulnikov, S.; Scheider, B. D.; Reyes, G. P.; Monfette, S.; Sigman, M. S. Hansen, E. C.; Weix, D. J. *J. Am. Chem. Soc.* **2024**, *146*, 3043–3051

Multivariate Linear Free Energy Relationships - Finding New Catalysts

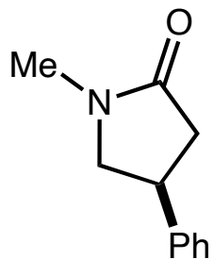


Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

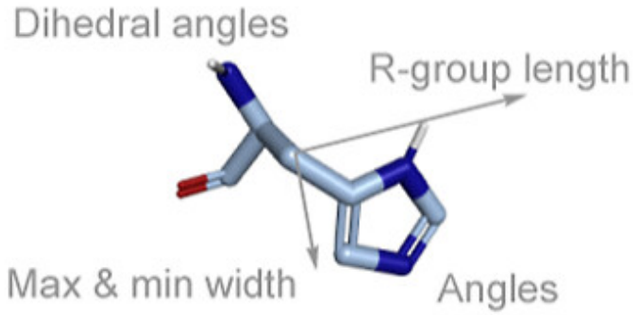
Multivariate Linear Free Energy Relationships - Finding New Catalysts



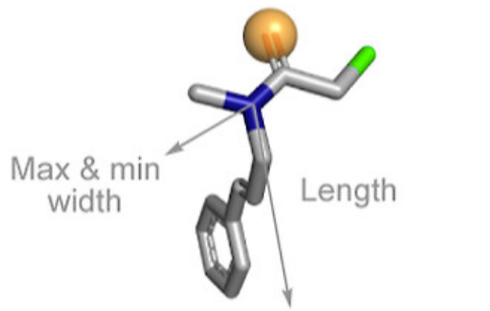
GluER-T36A 1%
NADP+ 1%
GDH-105, glucose
KP_i(100 mM, pH=8.0)
Cyan LEDs



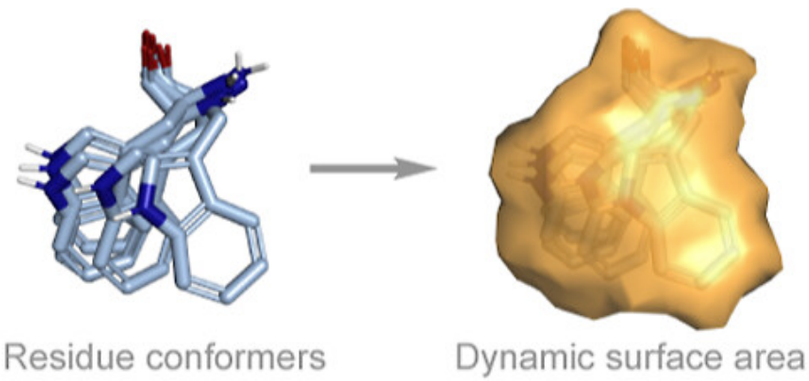
GluER-T36A



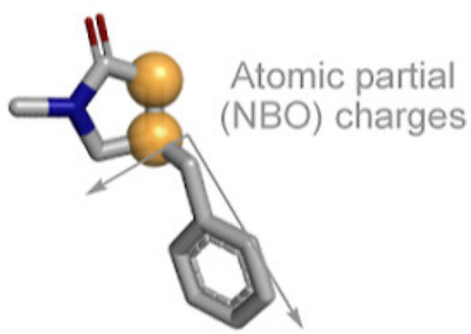
Geometric descriptors of relevant residues



Substrate sterics



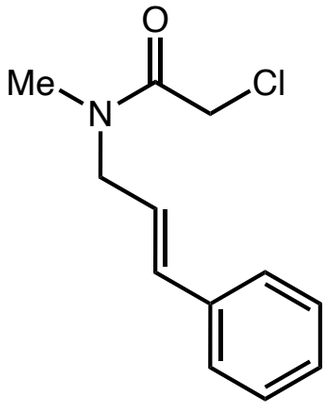
Enzyme dynamics



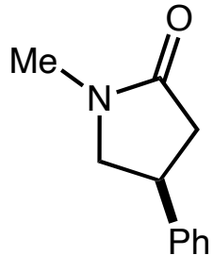
Substrate electronics

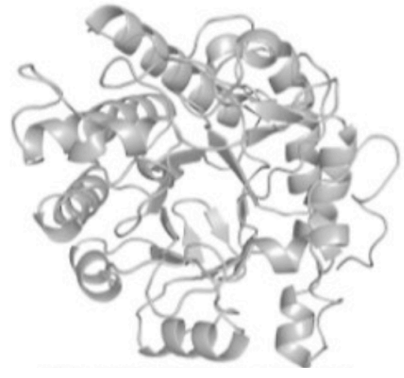
Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

Multivariate Linear Free Energy Relationships - Finding New Catalysts

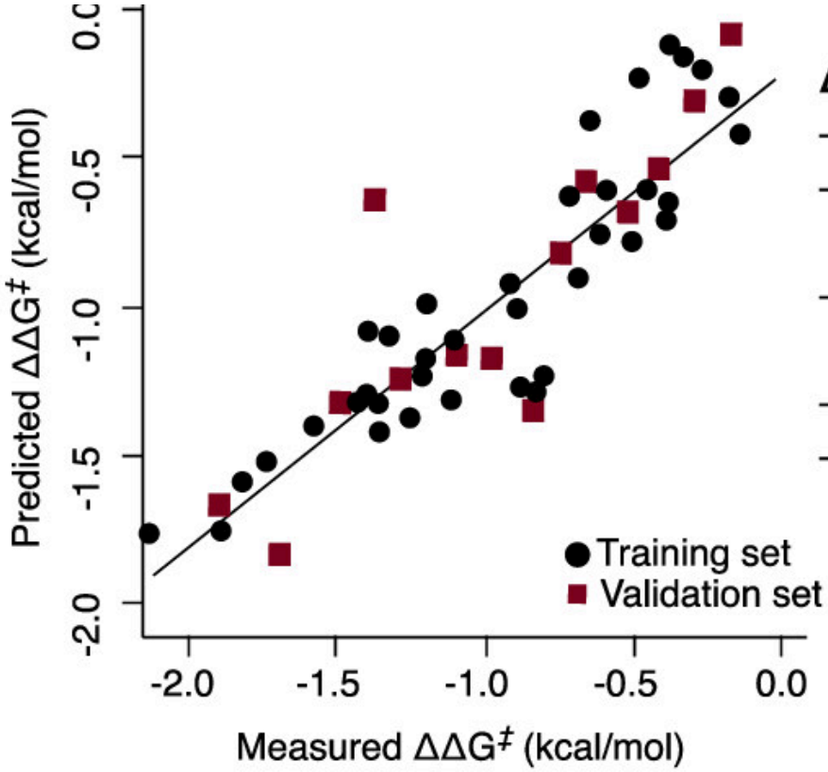


GluER-T36A 1%
NADP+ 1%
GDH-105, glucose
KP_i(100 mM, pH=8.0)
Cyan LEDs





GluER-T36A



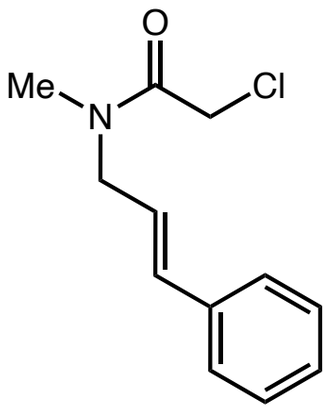
ΔΔG‡ Predicted =

– 0.96 +
– 0.12 * **ΔNBO$_{pdt\ ß\text{-}C}$**
0.09 * **ΔSterimol L $_{sub}$**
– 0.09 * **Residue 66$_{ΔSterimol\ L}$**
0.18 * **Residue 100$_{ΔAngle\ 1}$**
– 0.17 * **Residue 172$_{Sterimol\ L,\ max}$**
– 0.31 * **Residue 342$_{ΔAngle\ 3}$**

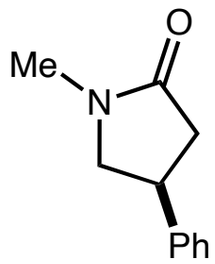
Training R^2 = 0.82
Validation R^2 = 0.73
Validation MAE = 0.19 kcal/mol

Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

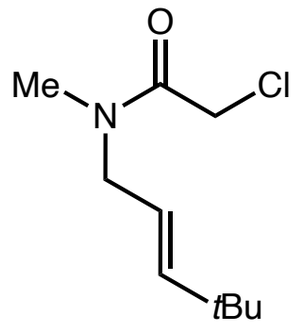
Multivariate Linear Free Energy Relationships - Finding New Catalysts



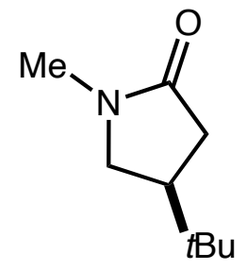
GluER-T36A 1%
NADP+ 1%
GDH-105, glucose
KP_i(100 mM, pH=8.0)
Cyan LEDs



GluER-T36A



Mutant prediction



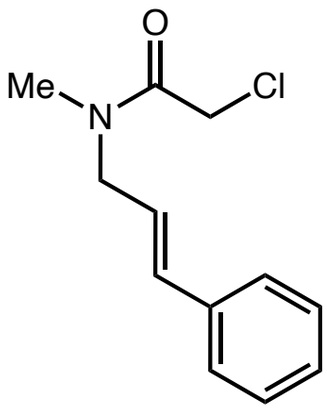
ΔΔG‡ Predicted =

– 0.96 +
– 0.12 * **ΔNBO$_{pdt\ ß\text{-}C}$**
0.09 * **ΔSterimol L $_{sub}$**
– 0.09 * **Residue 66$_{ΔSterimol\ L}$**
0.18 * **Residue 100$_{ΔAngle\ 1}$**
– 0.17 * **Residue 172$_{Sterimol\ L,\ max}$**
– 0.31 * **Residue 342$_{ΔAngle\ 3}$**

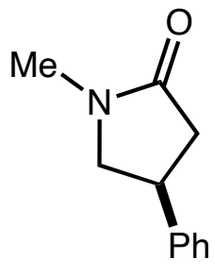
Training R^2 = 0.82
Validation R^2 = 0.73
Validation MAE = 0.19 kcal/mol

Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

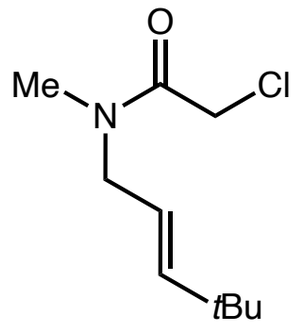
Multivariate Linear Free Energy Relationships - Finding New Catalysts



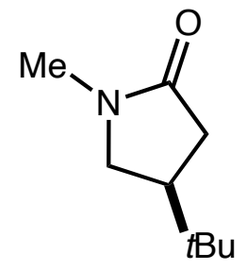
GluER-T36A 1%
NADP+ 1%
GDH-105, glucose
KP_i(100 mM, pH=8.0)
Cyan LEDs

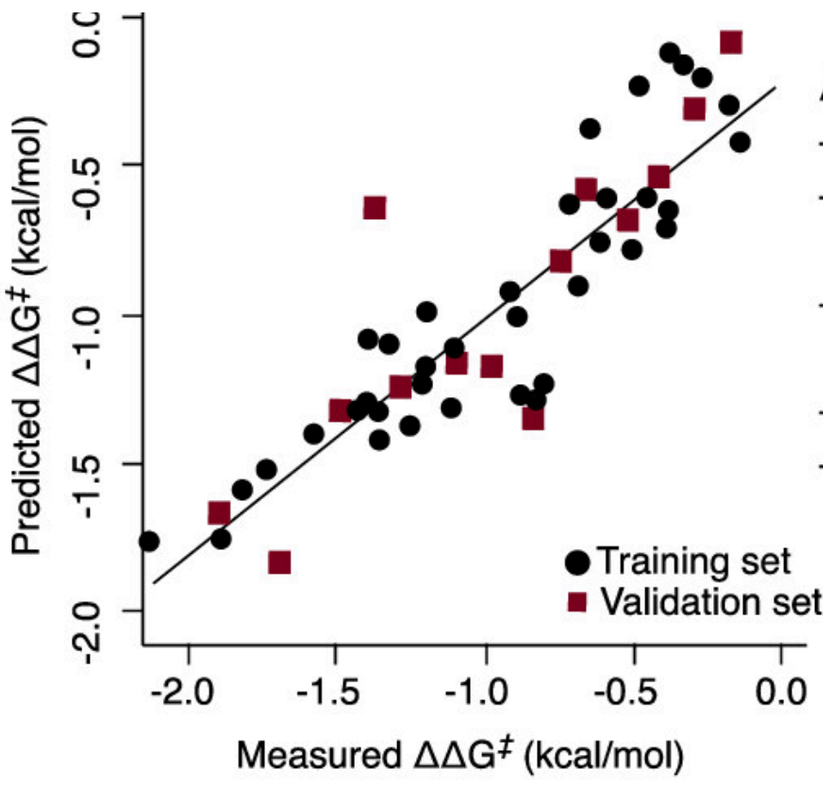


GluER-T36A



Mutant prediction





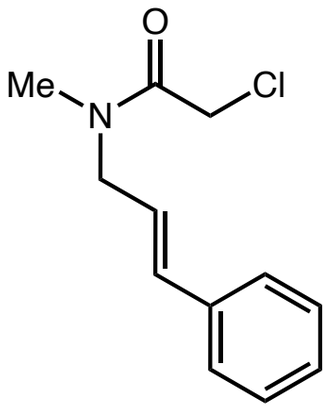
ΔΔG‡ Predicted =
– 0.96 +
– 0.12 * **ΔNBO$_{pdt\ ß\text{-}C}$**
0.09 * **ΔSterimol L $_{sub}$**
– 0.09 * **Residue 66$_{ΔSterimol\ L}$**
0.18 * **Residue 100$_{ΔAngle\ 1}$**
– 0.17 * **Residue 172$_{Sterimol\ L,\ max}$**
– 0.31 * **Residue 342$_{ΔAngle\ 3}$**

Training R^2 = 0.82
Validation R^2 = 0.73
Validation MAE = 0.19 kcal/mol

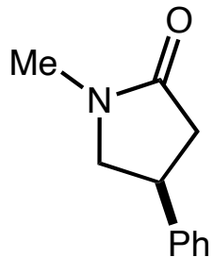
Entry	Pred. %ee[a]	Meas. %ee
5-W66A	34% ee	21% ee
5-W66L	34% ee	9% ee
5-Y177F	27% ee	36% ee
5-Q232F	44% ee	3% ee
5-Y343A	23% ee	29% ee
5-Y343F	80% ee	72% ee
5-Y343W	44% ee	38% ee

Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

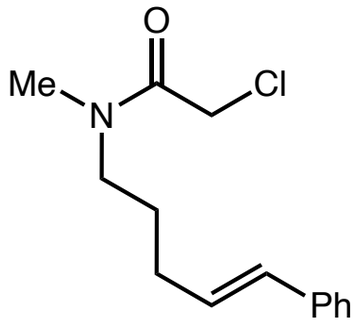
Multivariate Linear Free Energy Relationships - Finding New Catalysts



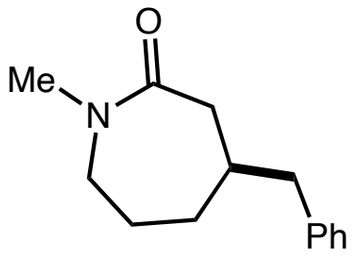
GluER-T36A 1%
NADP+ 1%
GDH-105, glucose
KP_i(100 mM, pH=8.0)
Cyan LEDs



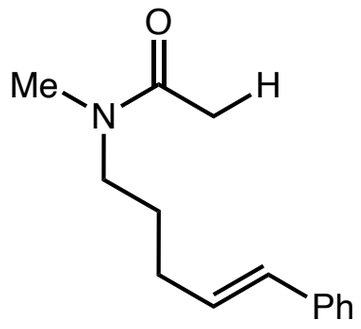
GluER-T36A



Mutant prediction



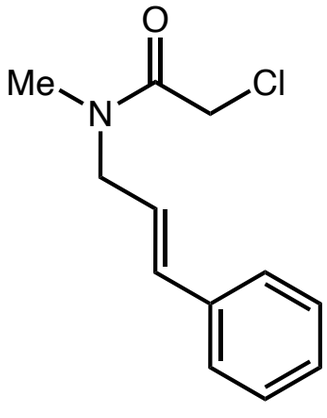
6b



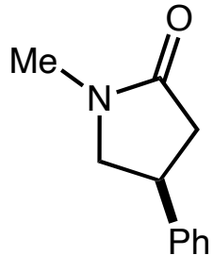
6c

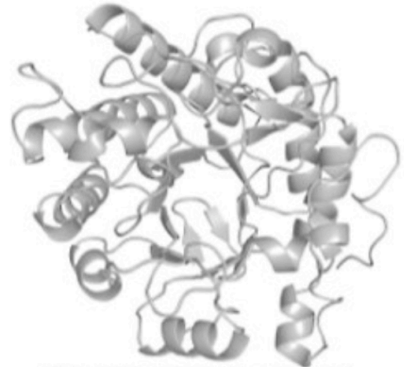
Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

Multivariate Linear Free Energy Relationships - Finding New Catalysts

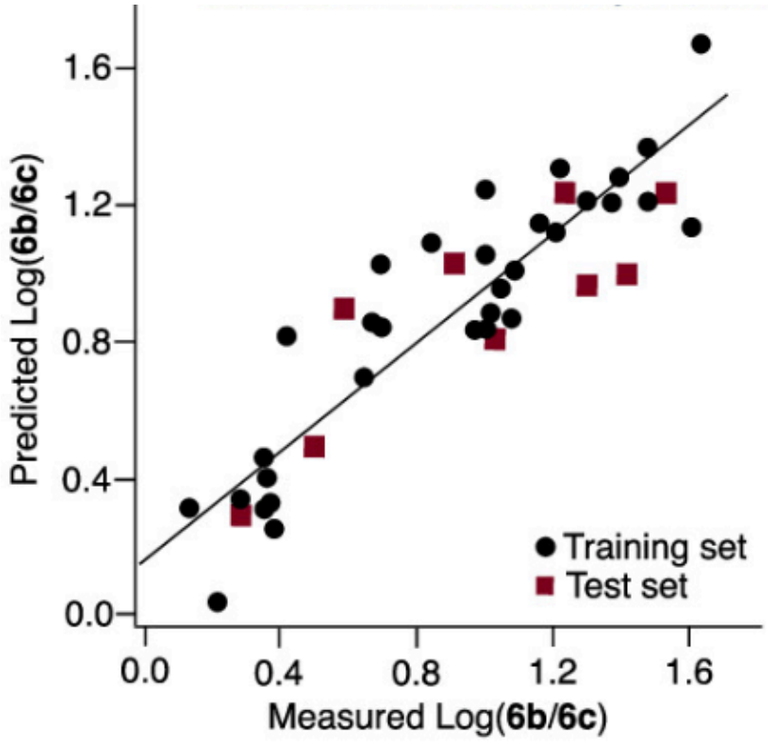


GluER-T36A 1%
NADP+ 1%
GDH-105, glucose
KP_i(100 mM, pH=8.0)
Cyan LEDs





GluER-T36A



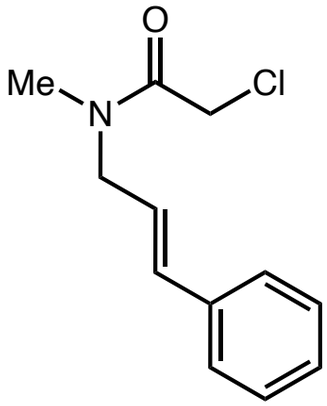
$\Delta\Delta G^{\ddagger}$ Predicted =

0.90 +
– 0.33 * **$NBO_{pdt\ carbonyl\ C,\ GS}$**
– 0.23 * **$NBO_{pdt\ ß\text{-}H,\ GS}$**
– 0.29 * **$Sterimol\ L_{pdt,\ min}$**
0.13 * **$Residue\ 269_{Sterimol\ B5\ sub,\ GS}$**

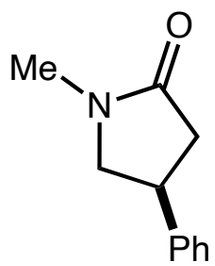
Training R^2 = 0.82
Validation R^2 = 0.70
Validation MAE = 0.19 kcal/mol

Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

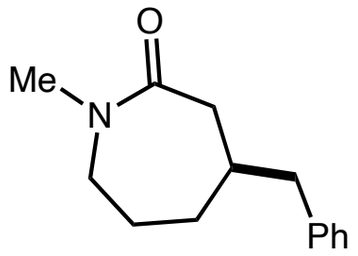
Multivariate Linear Free Energy Relationships - Finding New Catalysts



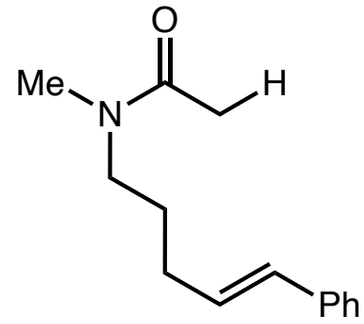
GluER-T36A 1%
NADP+ 1%
GDH-105, glucose
KP_i(100 mM, pH=8.0)
Cyan LEDs



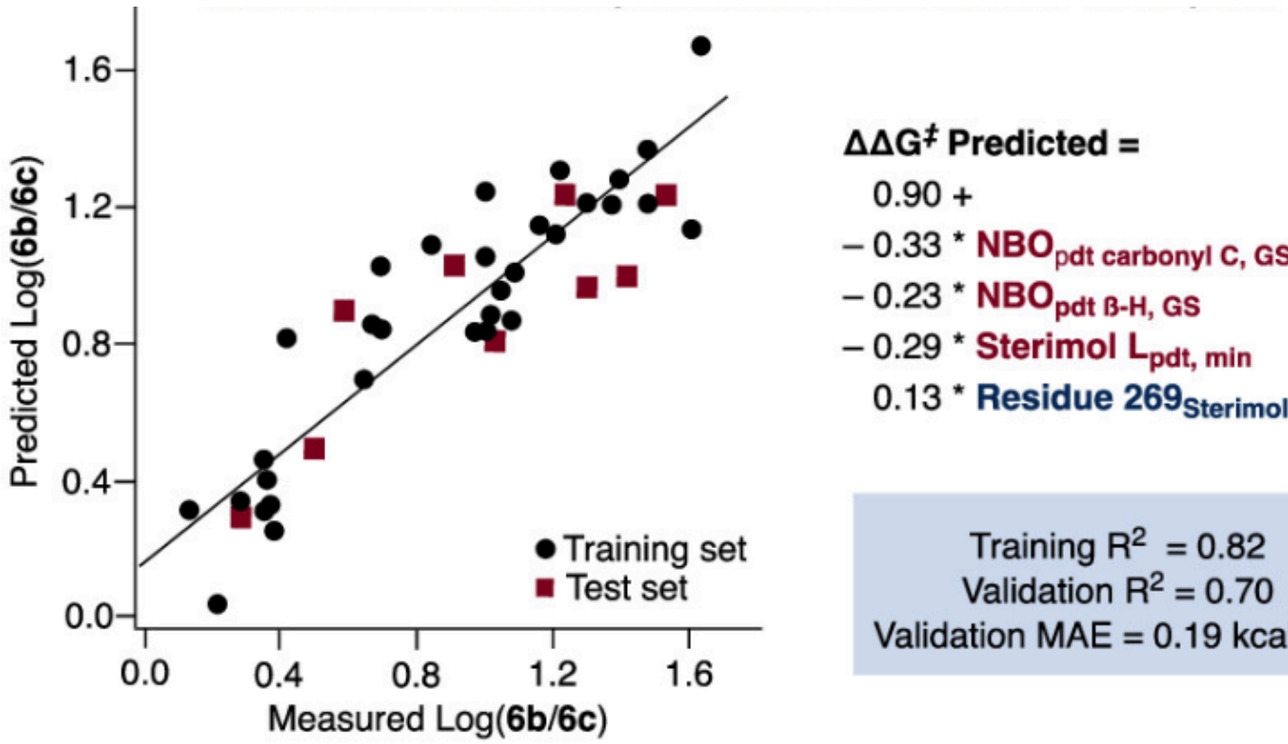
GluER-T36A



6b



6c



ΔΔG‡ Predicted =

0.90 +
– 0.33 * $NBO_{pdt\ carbonyl\ C,\ GS}$
– 0.23 * $NBO_{pdt\ ß\text{-}H,\ GS}$
– 0.29 * $Sterimol\ L_{pdt,\ min}$
0.13 * $Residue\ 269_{Sterimol\ B5\ sub,\ GS}$

Training R^2 = 0.82
Validation R^2 = 0.70
Validation MAE = 0.19 kcal/mol

Entry	Pred. (6b/6c)	Meas. (6b/6c)
6-W66A	4.0	4.6
6-W66L	3.9	9.2
6-Y177F	1.9	7.9
6-Q232F	2.8	4.3
6-Y343A	4.7	6.6
6-Y343F	2.7	2.8
6-Y343W	2.6	n.r.

Clements, H. D.; Flynn, A. R.; Nicholls, B. T.; Grosheva, D.; Leface, S. J.; Merriman, M. T.; Hyster, T. K.; Sigman, M. S.*J. Am. Chem. Soc.* **2023**, *145*, 17656–17664

Additional Reading

Data Science in Chemistry

Reviews

Williams, W. L. Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G., **Anslyn, E. V.** *ACS Cent. Sci.* **2021**, *21*, 1622–1637

Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. *ACS Cent. Sci.* **2023**, *9*, 2196–2204

Crawford, J. M.; Kingston, C.; **Toste, D. F**.; Sigman, M. S. *Acc. Chem. Res.* **2021**, *54*, 3136–3148

Additional Examples

Morak, T.; Myers, T. E.; Karas, L. J.; Hardy, M. A.; Mercado, B. Q.; Sigman, M. S.; **Miller, S. J.** *J. Am. Chem. Soc.* **2023**, *145*, 22322–22328

Nistanaki, S. K.; Williams, C. G.; Wigman, B.; Wong, J. J.; Haas, B. C.; Popov, S.; Werth, J.; Sigman, M. S.; Houk, K. N.; **Nelson, H. M**. *Science*, **2022**, *378*, 1085–1091

Sasha, M. H.; Wahlman, J. L. H.; Read, J. A.; Werth, J.; **Jacobsen, E. N**.; Sigman, M. S. *ACS Catal.* **2022**, *12*, 14836–14845

Boni, Y. T.; Cammarota, R. C.; Liao, K.; Sigman, M. S.; **Davies, H. M. L**. *J. Am. Chem. Soc.* **2022**, *144*, 15549–15591

Modern approaches to methods development

Modern Paradigms in Screening

- *Reaction generalization*
- *“Accelerate” Serendipity*
- *Miniaturization of unique reaction set ups*

Data Science

- *Catalyst optimization*
- *Predicting selectivity*
- *Discovery of new catalysts*

Machine Learning

- *What is machine learning*
- *Prediction of optimal conditions*
- *Selectivity prediction for complex systems*
- *Catalyst Discovery*

What is Machine Learning?

General use of algorithms and data to create autonomous or semi–autonomous tasks

Define a task

identify apples?

What is Machine Learning?

General use of algorithms and data to create autonomous or semi–autonomous tasks

Define a task

identify apples?

Parameterize Data

Pictures of apples

What is Machine Learning?

General use of algorithms and data to create autonomous or semi–autonomous tasks

Define a task

identify apples?

Parameterize Data

Pictures of apples

Curate Data

Is an object an apple?

What is Machine Learning?

General use of algorithms and data to create autonomous or semi–autonomous tasks

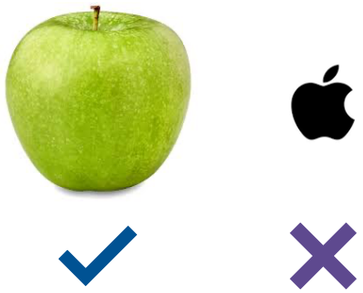
Define a task

identify apples?

Parameterize Data

Pictures of apples

Curate Data



Is an object an apple?

Choose Model

Train the AI

What is Machine Learning?

General use of algorithms and data to create autonomous or semi–autonomous tasks

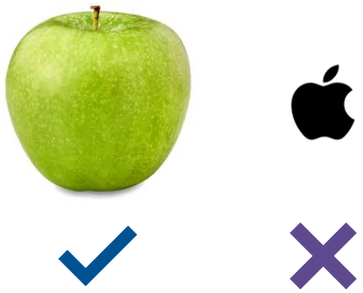
Define a task

identify apples?

Parameterize Data

Pictures of apples

Curate Data



Is an object an apple?

Choose Model

Train the AI

Validate Model

Input

Apples

Output

What is Machine Learning?

General use of algorithms and data to create autonomous or semi–autonomous tasks

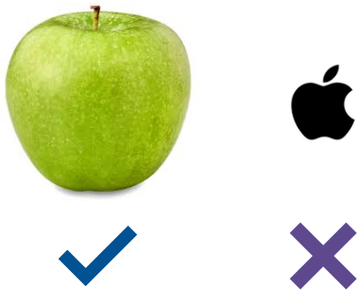
Define a task

identify apples?

Parameterize Data

Pictures of apples

Curate Data



Is an object an apple?

Choose Model

Train the AI

Application

Validate Model

Input

Apples

Output

What is Machine Learning?

General use of algorithms and data to create autonomous or semi–autonomous tasks

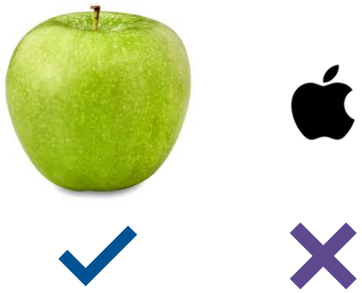
Define a task

identify apples?

Parameterize Data

Pictures of apples

Curate Data



Is an object an apple?

Choose Model

Train the AI

Application

Same algorithm to search through as many pictures as you want

Validate Model

Input

Apples

Output

Machine Learning Applied to Chemistry

General use of algorithms and data to create autonomous or semi–autonomous tasks

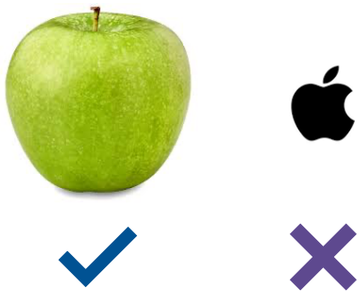
Define a task

identify apples?

Parameterize Data

Pictures of apples

Curate Data



Is an object an apple?

Choose Model

Train the AI

Application

Same algorithm to search through as many pictures as you want

Validate Model

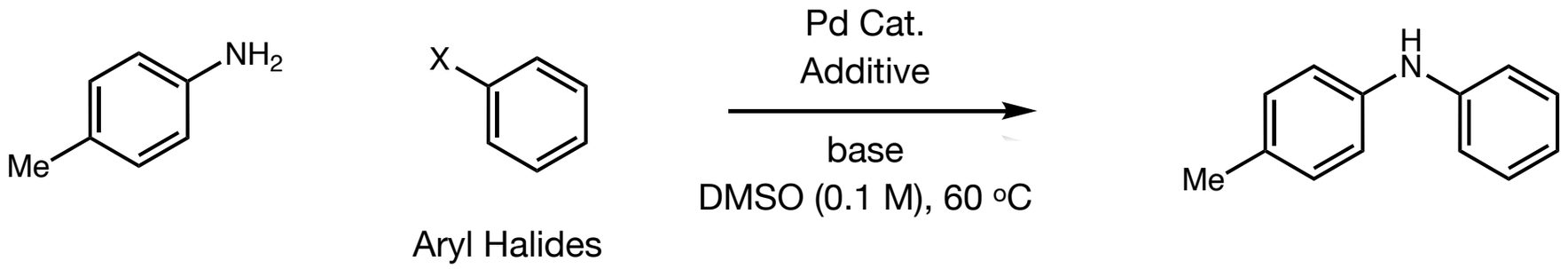
Input

Apples

Output

Machine Learning Applied to Chemistry

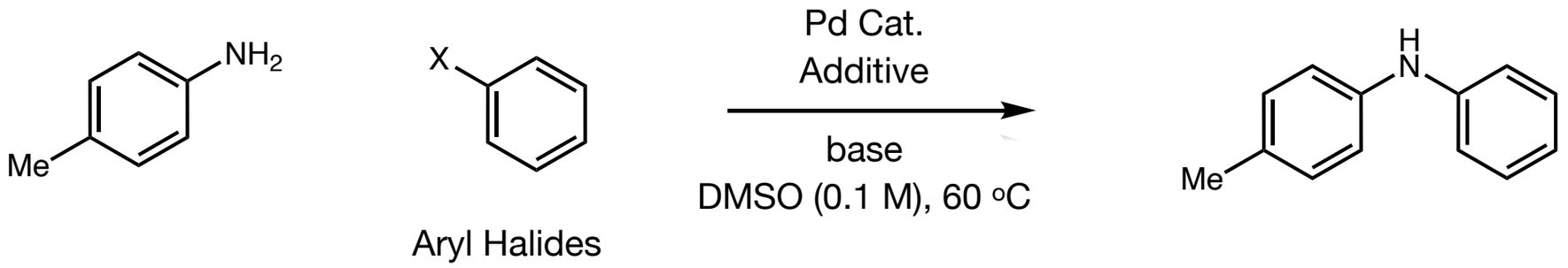
Task: *Predict the yield of the Buchwald-Hartwig reaction under varying reaction conditions*



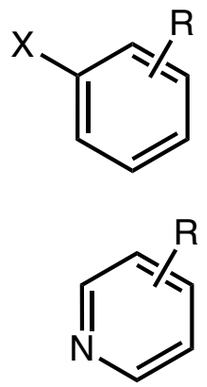
Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry

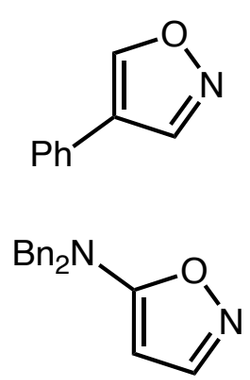
Task: *Predict the yield of the Buchwald-Hartwig reaction under varying reaction conditions*



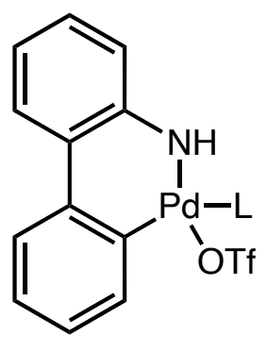
Variables



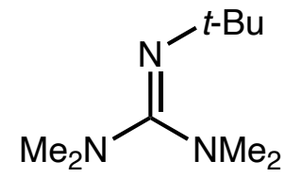
Aryl Halides (**15**)



Additives (**23**)



Pd Cat. (**4**)

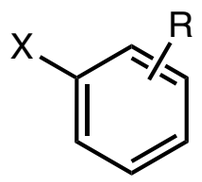


Bases (**3**)

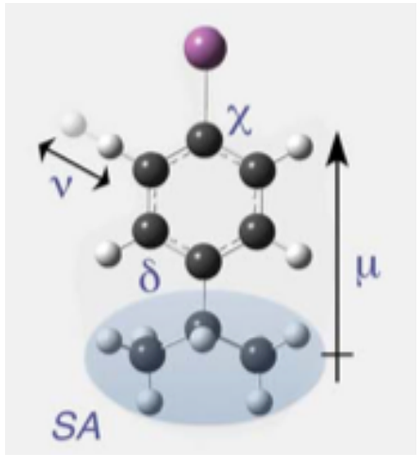
How can we feed data to the algorithm about each variable?

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry



Spartan (software)



Parameterized Data

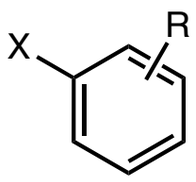
- *Vibrational properties*
- *Surface area*
- *Dipole moment*
- *Atomic electrostatic charge*

Total of 120 descriptors obtained for each individual reaction

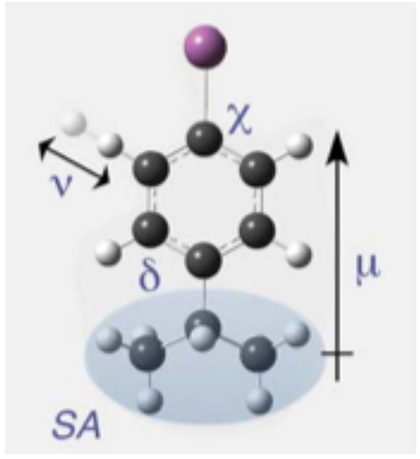
What is the data input?

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry



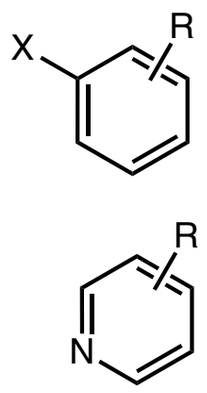
Spartan (software)



Parameterized Data

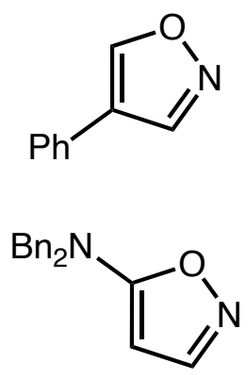
- *Vibrational properties*
- *Surface area*
- *Dipole moment*
- *Atomic electrostatic charge*

Total of 120 descriptors obtained for each individual reaction



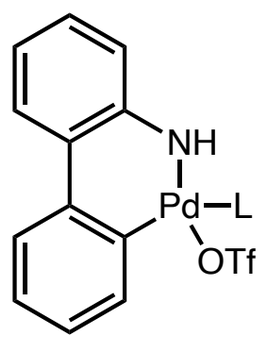
Aryl Halides (**15**)

×



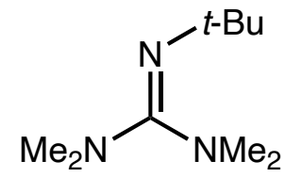
Additives (**23**)

×



Pd Cat. (**4**)

×

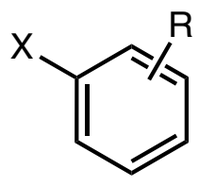


Bases (**3**)

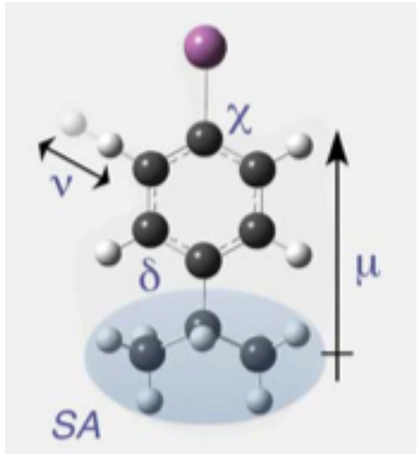
Run 4140 reactions through HTE and input yield along with parameterized reaction conditions

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry



Spartan (software)



Parameterized Data

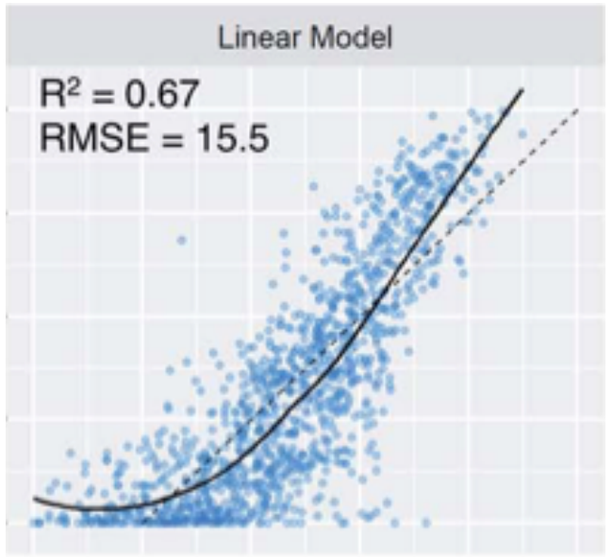
- *Vibrational properties*
- *Surface area*
- *Dipole moment*
- *Atomic electrostatic charge*

Total of 120 descriptors obtained for each individual reaction

Unlike MLFER all 120 descriptors can be used in the same model in a null-hypothesis manner

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry



RMSE: root mean square error

Accuracy

R^2: Coefficient of determination

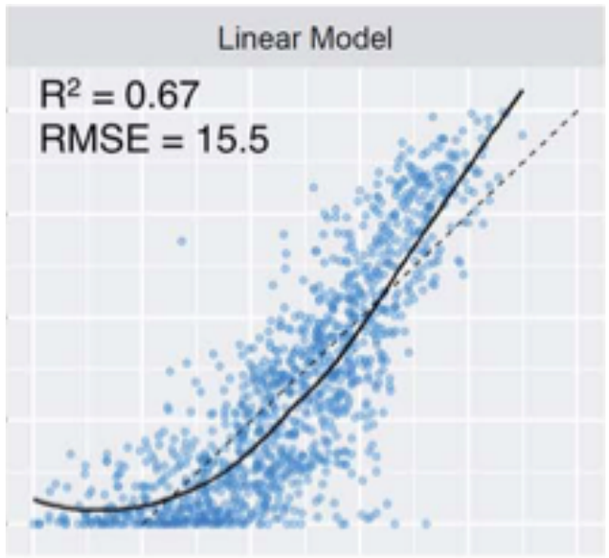
Consistency

straight, well correlated fits are more accurate models

Linear Regression Model is inaccurate and inconsistent

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry



RMSE: root mean square error

↓

Accuracy

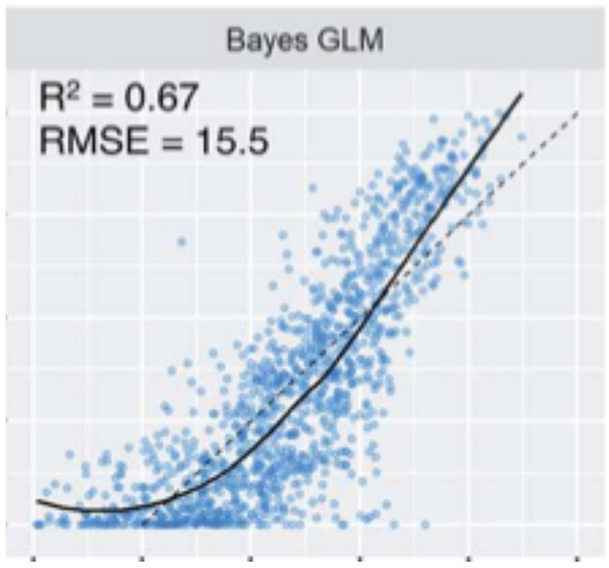
R^2: Coefficient of determination

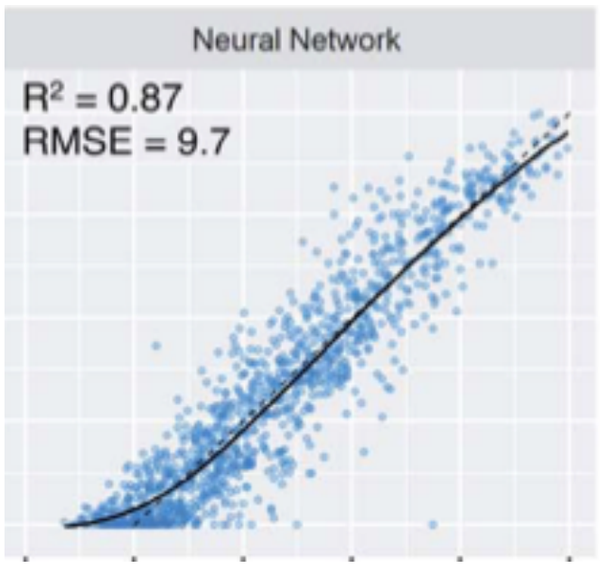
↓

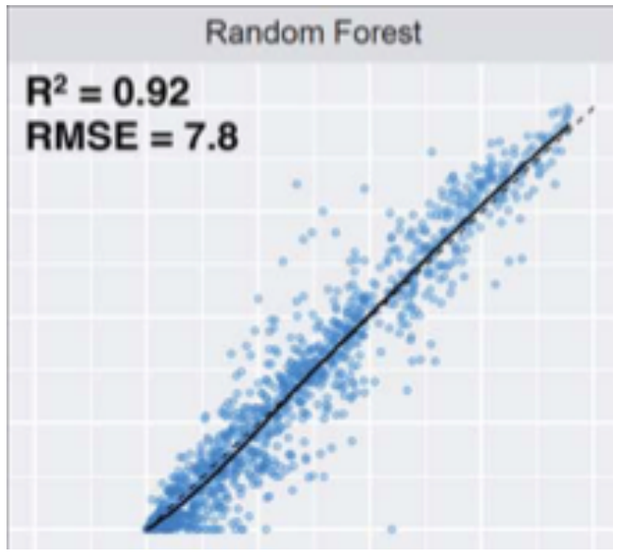
Consistency

straight, well correlated fits are more accurate models

supervised machine learning models

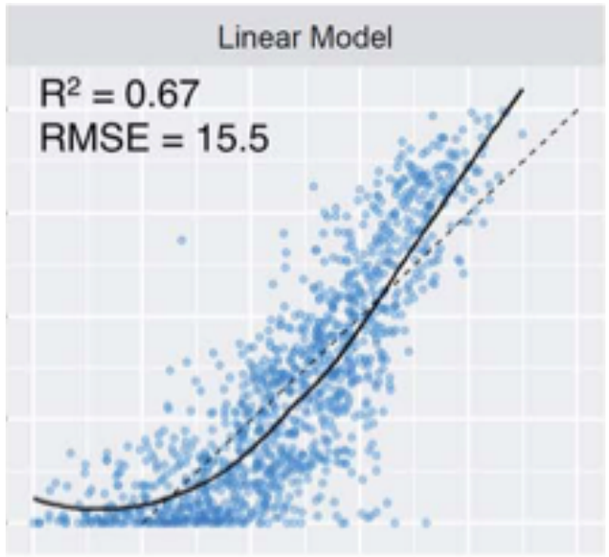






Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry

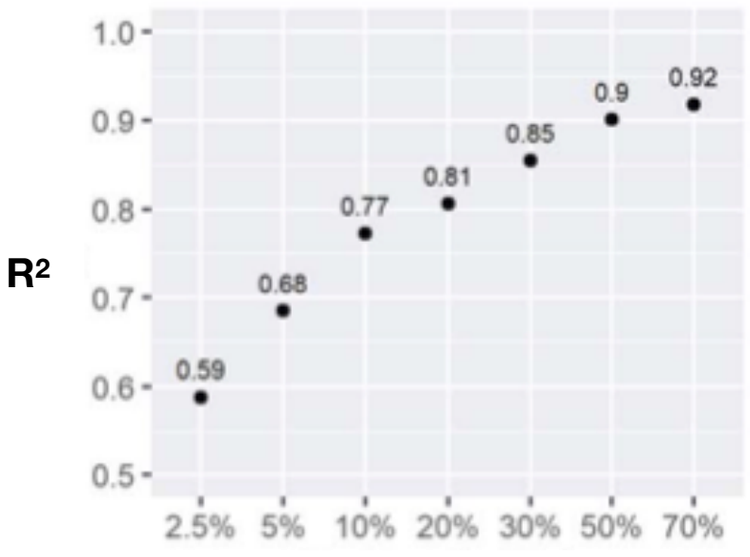


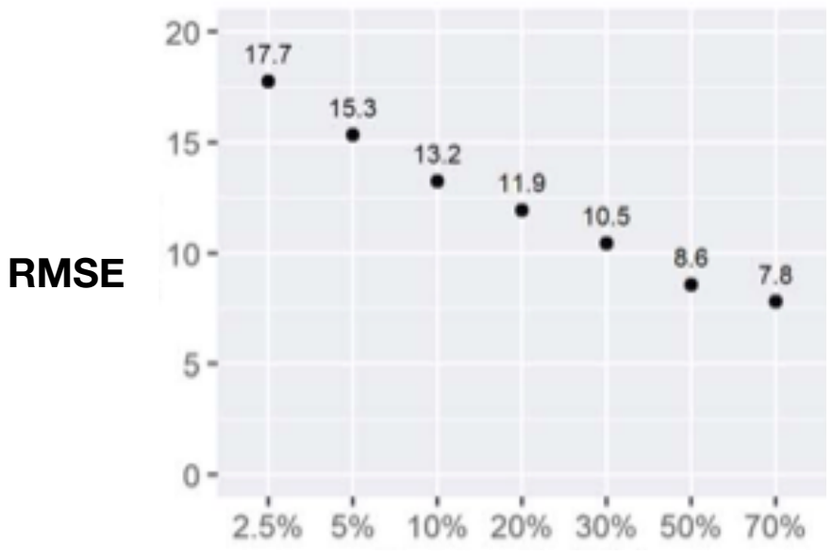
RMSE: root mean square error → **Accuracy**

R^2: Coefficient of determination → **Consistency**

straight, well correlated fits are more accurate models

supervised machine learning models





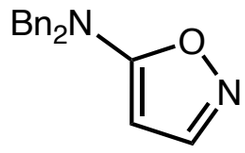
Bigger training set makes predictions more consistent and more accurate

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

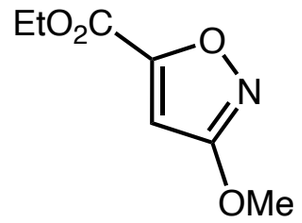
Machine Learning Applied to Chemistry

70% of data used to train model

30% of data used to validate model

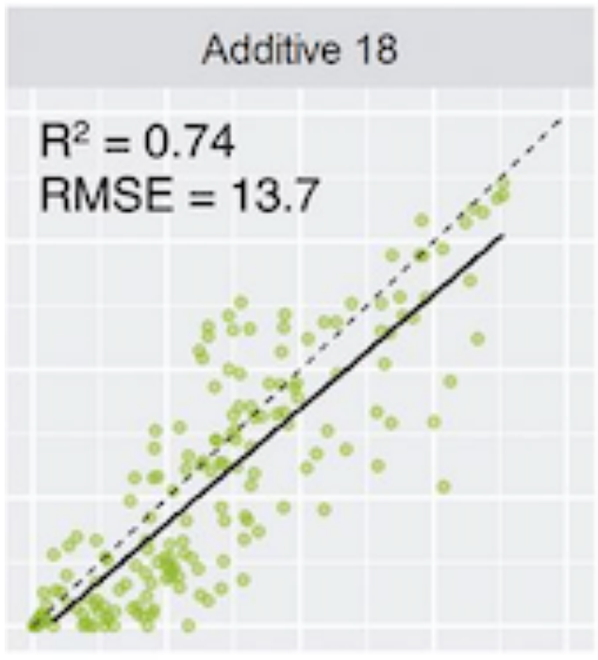


Additive 18

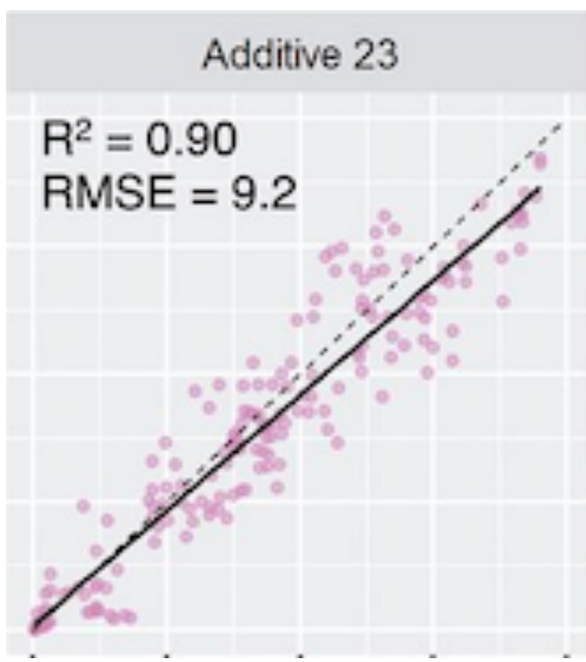


Additive 23

Poorly predictive



predictive



observed yield

predicted yield

predicted yield

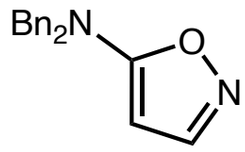
What causes the difference in accuracy between additive 18 and 23?

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

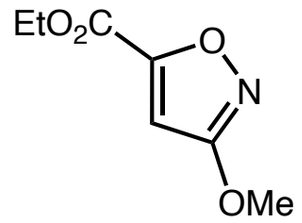
Machine Learning Applied to Chemistry

70% of data used to train model

30% of data used to validate model

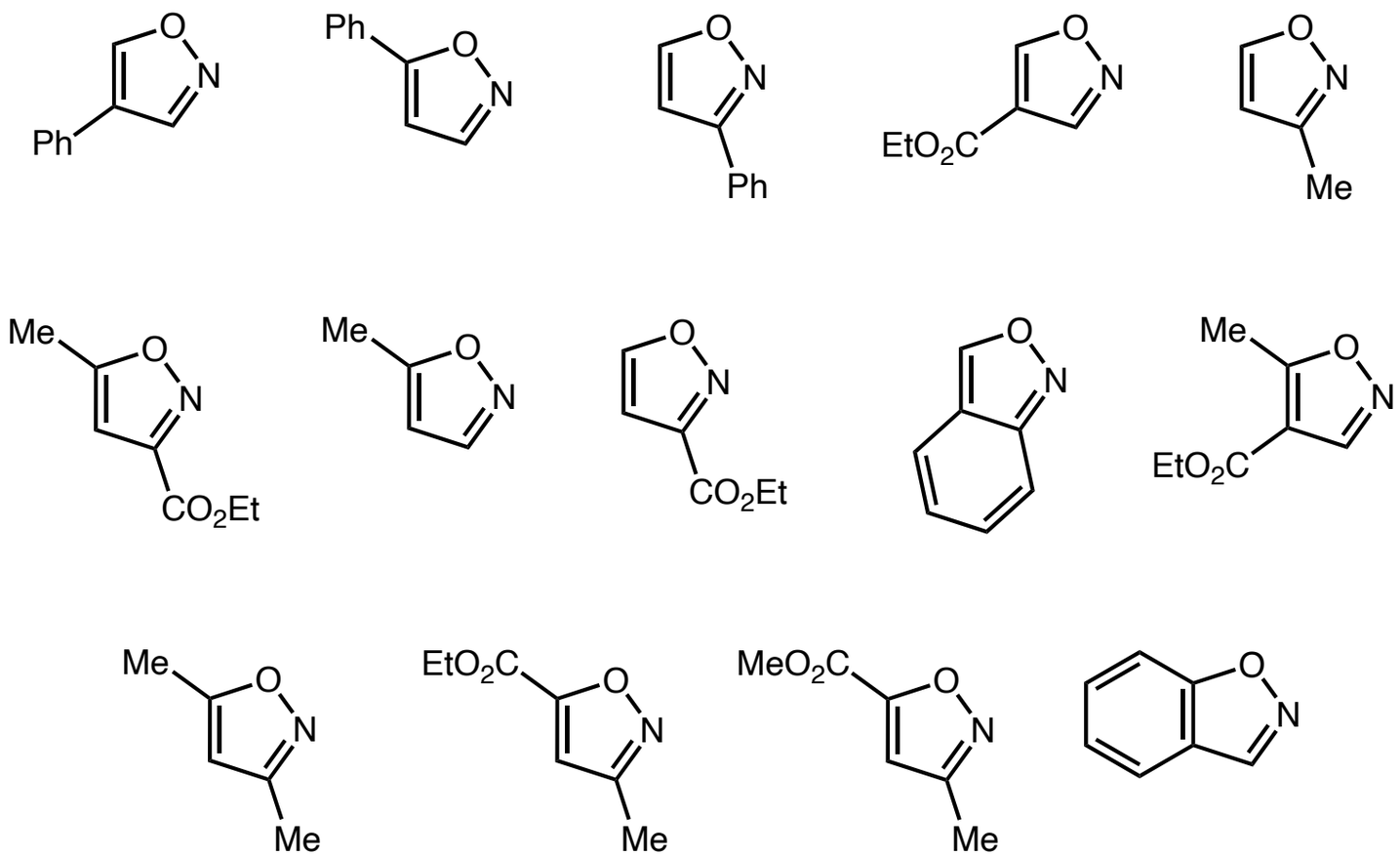


Additive 18



Additive 23

training set additives



All “electron poor” to neutral isoxazoles

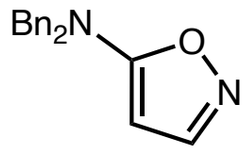
Model poorly predicts yields for additive 18 because it falls out of the scope of the training set

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

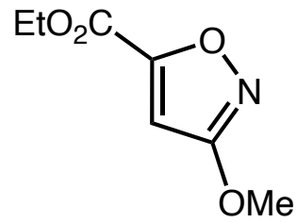
Machine Learning Applied to Chemistry

70% of data used to train model

70% of data used to train model

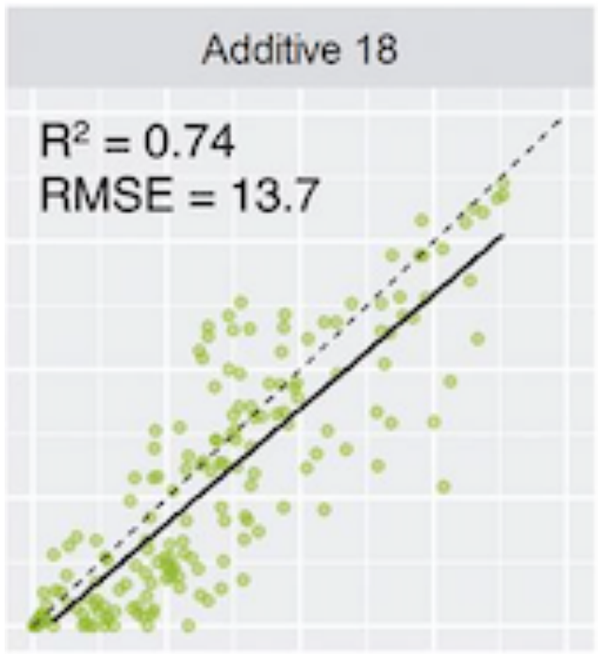


Additive 18



Additive 23

Poorly predictive

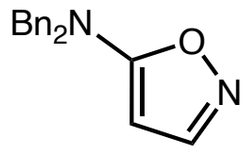


Inaccurate for parameters outside of training data set

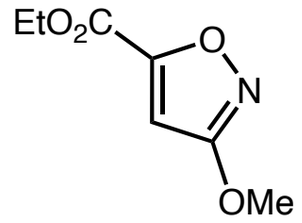
Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry

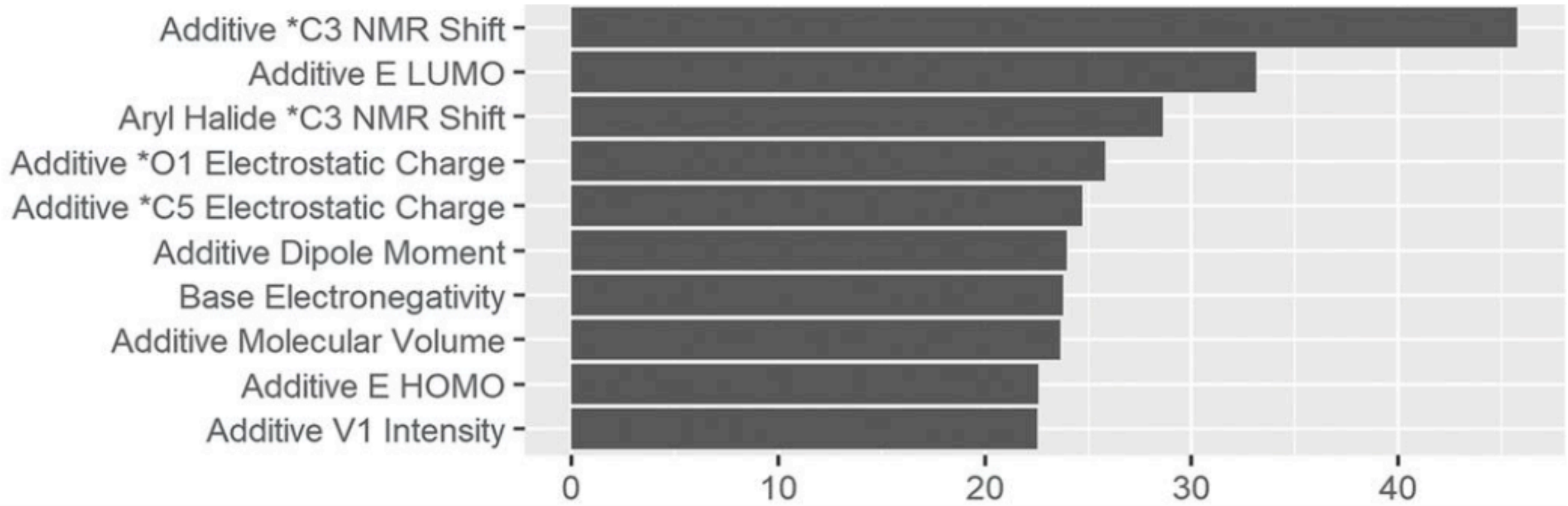
Why does the identity of the additive matter?



Additive 18



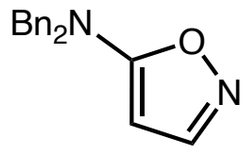
Additive 23



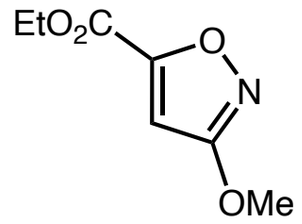
Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry

Why does the identity of the additive matter?



Additive 18

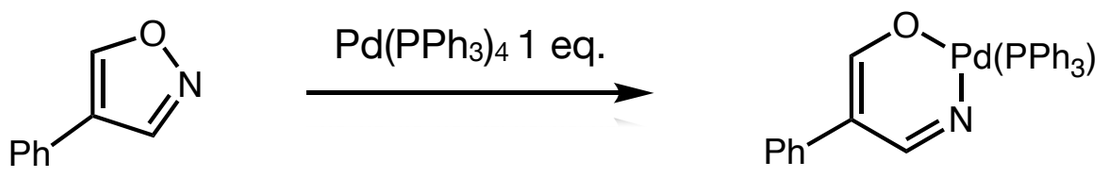


Additive 23

Descriptor

Additive *C3 NMR Shift
Additive E LUMO
Aryl Halide *C3 NMR Shift
Additive *O1 Electrostatic Charge
Additive *C5 Electrostatic Charge
Additive Dipole Moment
Base Electronegativity
Additive Molecular Volume
Additive E HOMO
Additive V1 Intensity

Top descriptors suggest additive electrophilicity influences yield

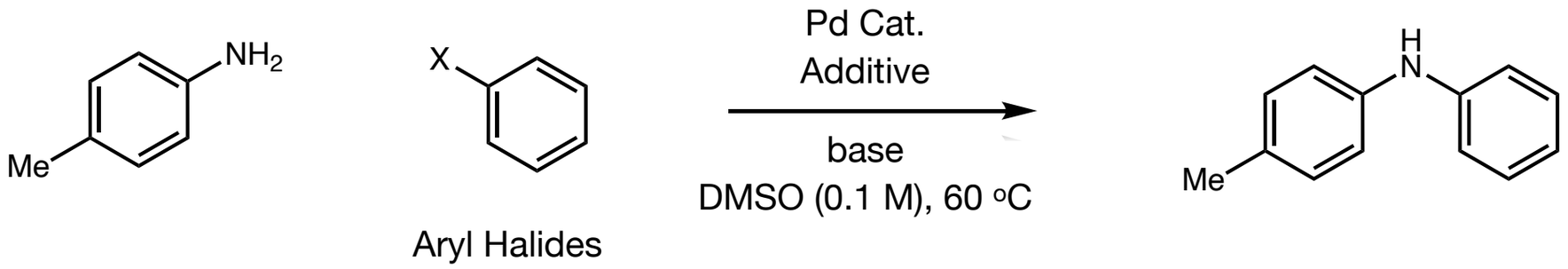


known for nickel but not for palladium

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry

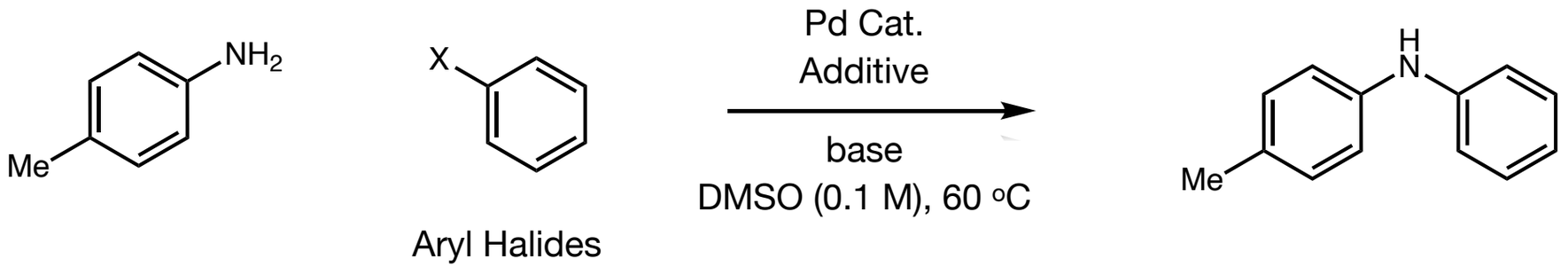
Task: *Predict the yield of the Buchwald-Hartwig reaction under varying reaction conditions*



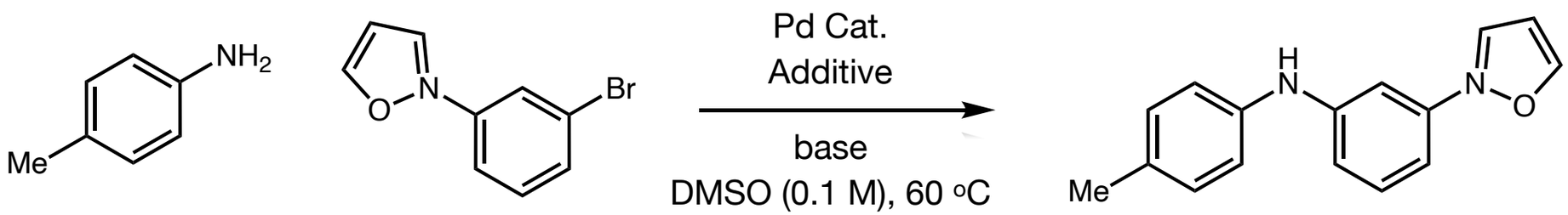
Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry

Task: *Predict the yield of the Buchwald-Hartwig reaction under varying reaction conditions*



Result: *Model that can roughly predict the yield of a Buchwald Hartwig reaction in which isoxazoles are present*



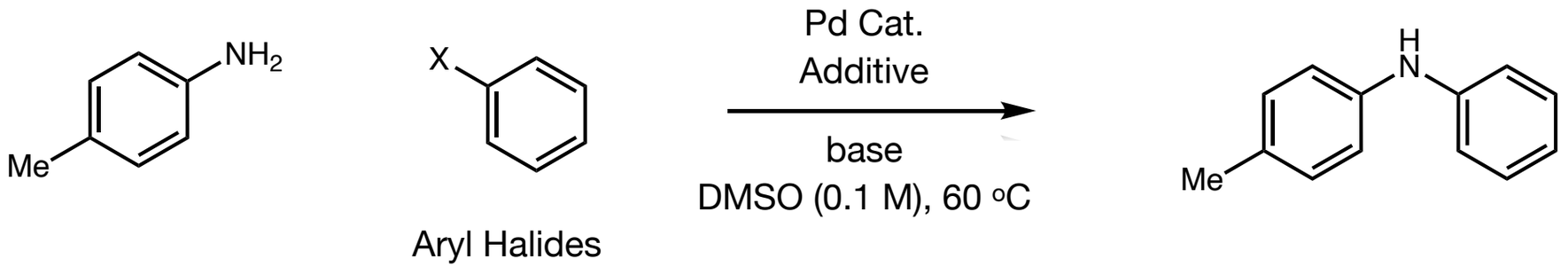
relatively specific task

Needed 4,140 experimental yields

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry

Task: *Predict the yield of the Buchwald-Hartwig reaction under varying reaction conditions*



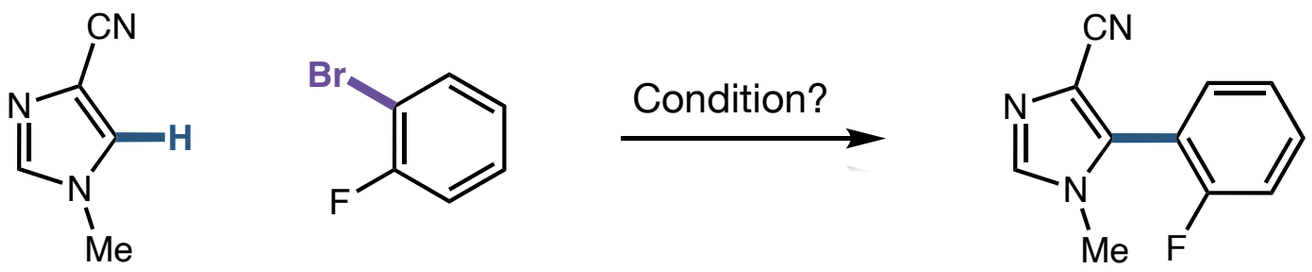
Result: *Model that can roughly predict the yield of a Buchwald Hartwig reaction in which isoxazoles are present*

strategy not fit for reaction optimization

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Doyle, A. G. *Science* **2018**, 360, 186–190.

Machine Learning Applied to Chemistry - Reaction Optimization

Task: *Optimize a reaction where 1000's of experimental yields are not available*

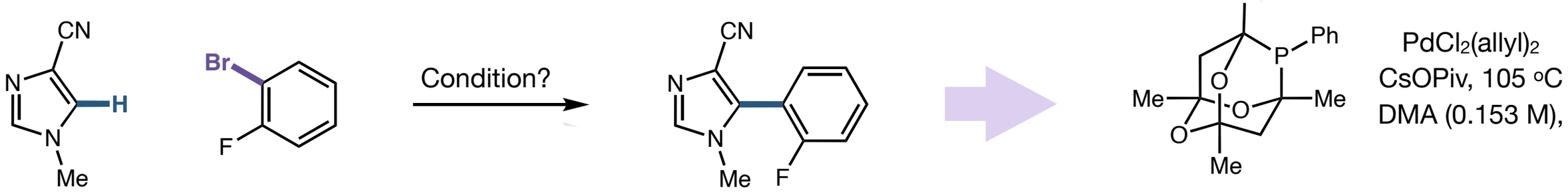


No previous data

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

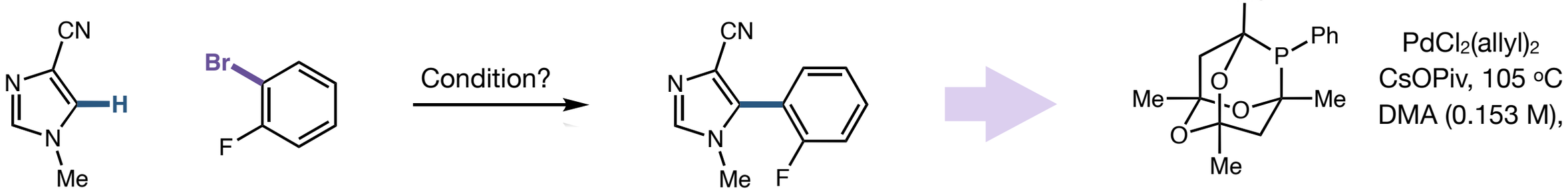
Task: *Optimize a reaction where 1000's of experimental yields are not available*

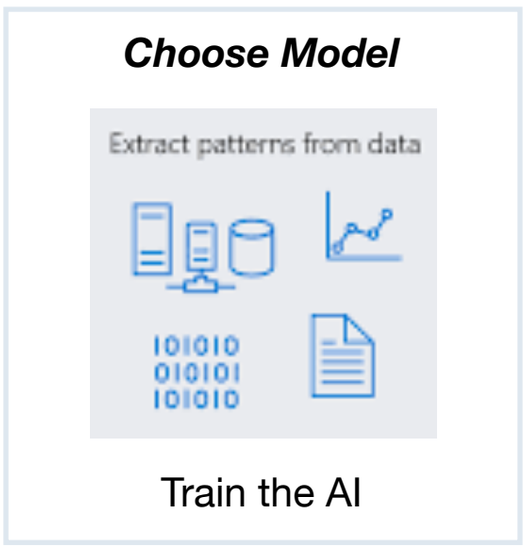


Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Task: *Optimize a reaction where 1000's of experimental yields are not available*





How would you train a model that trains itself for new systems?

What would the model need to do to quickly find optimal conditions?

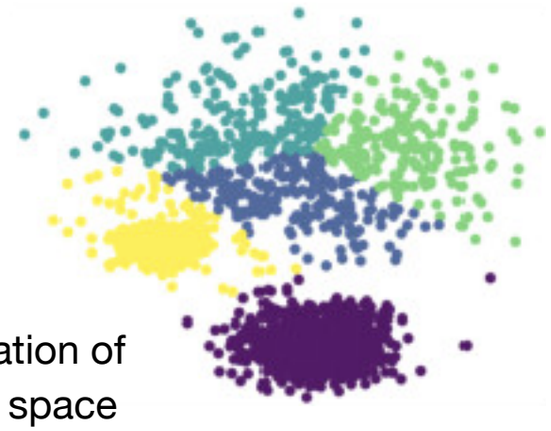
Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Choose Model

Train the AI

1. Model potential condition space



2D representation of condition space

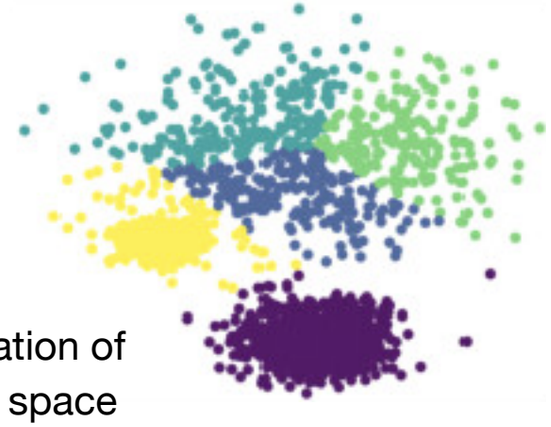
Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Choose Model

Train the AI

1. Model potential condition space



2D representation of condition space

2. Choose a set of conditions to evaluate

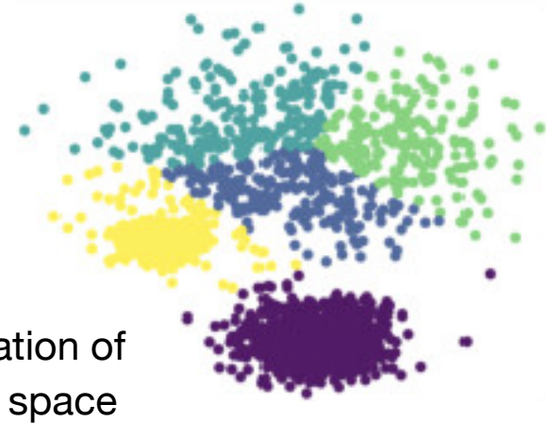
Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Choose Model

Train the AI

1. Model potential condition space



2D representation of condition space

2. Choose a set of conditions to evaluate

3. Human runs experiments and inputs data

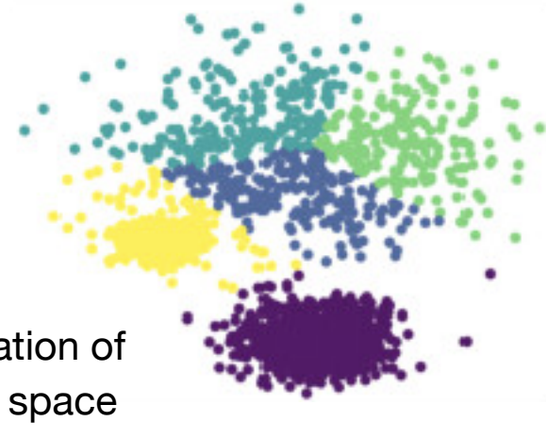
Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Choose Model

Train the AI

1. Model potential condition space



2D representation of condition space

2. Choose a set of conditions to evaluate

3. Human runs experiments and inputs data

4. Redefine better set of conditions to evaluate

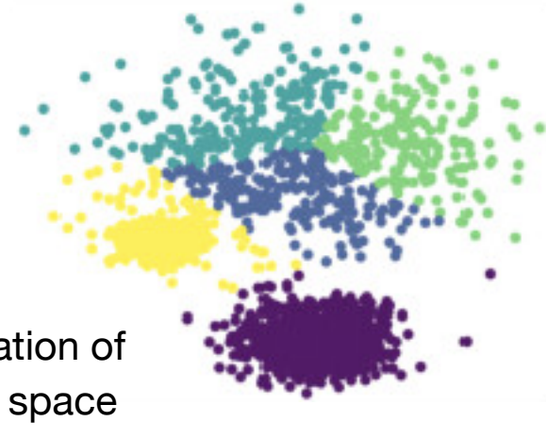
Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Choose Model

Train the AI

1. Model potential condition space



2D representation of condition space

2. Choose a set of conditions to evaluate

3. Human runs experiments and inputs data

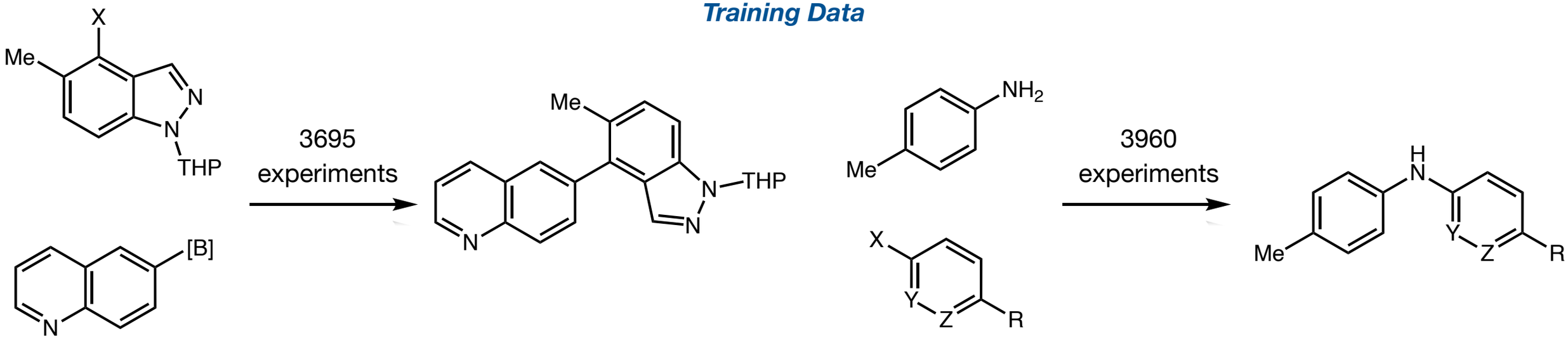
4. Redefine better set of conditions to evaluate

5. Repeat until yield optimized or experimental resources run out

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

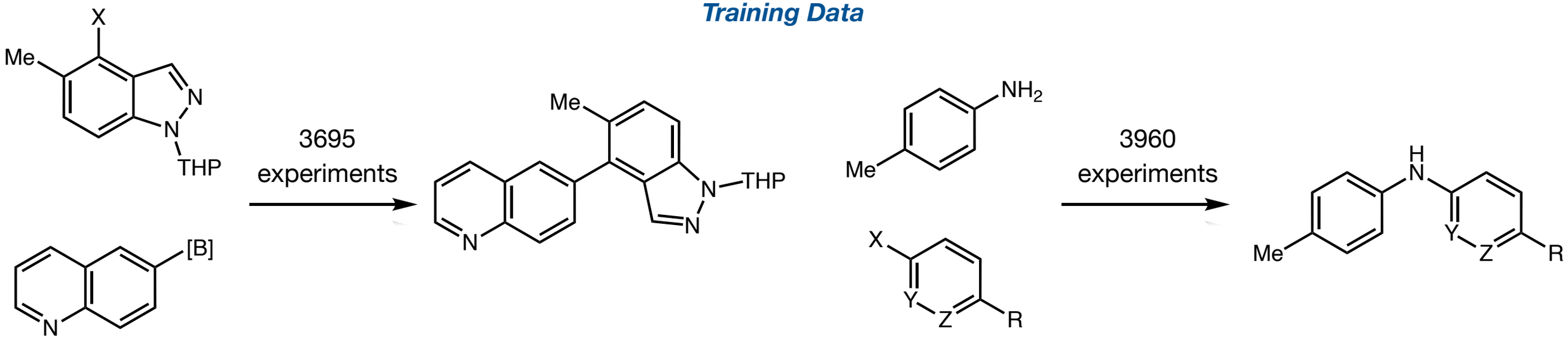
Training Data

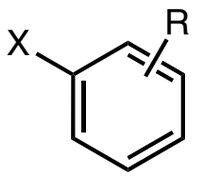


Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

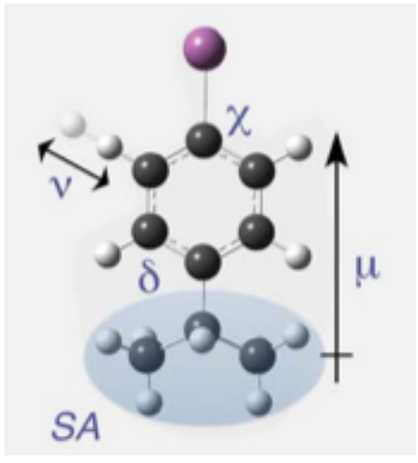
Machine Learning Applied to Chemistry - Reaction Optimization

Training Data





Automated DFT

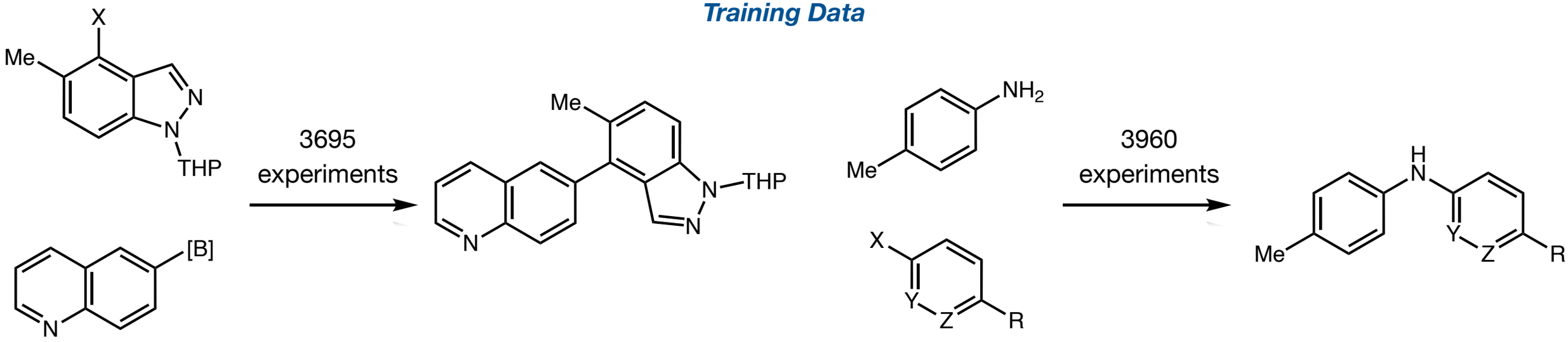


Just need to input SMILES

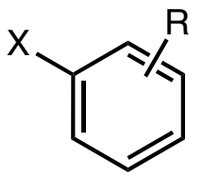
Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

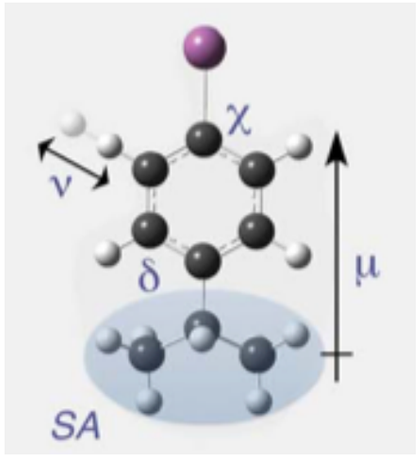
Training Data

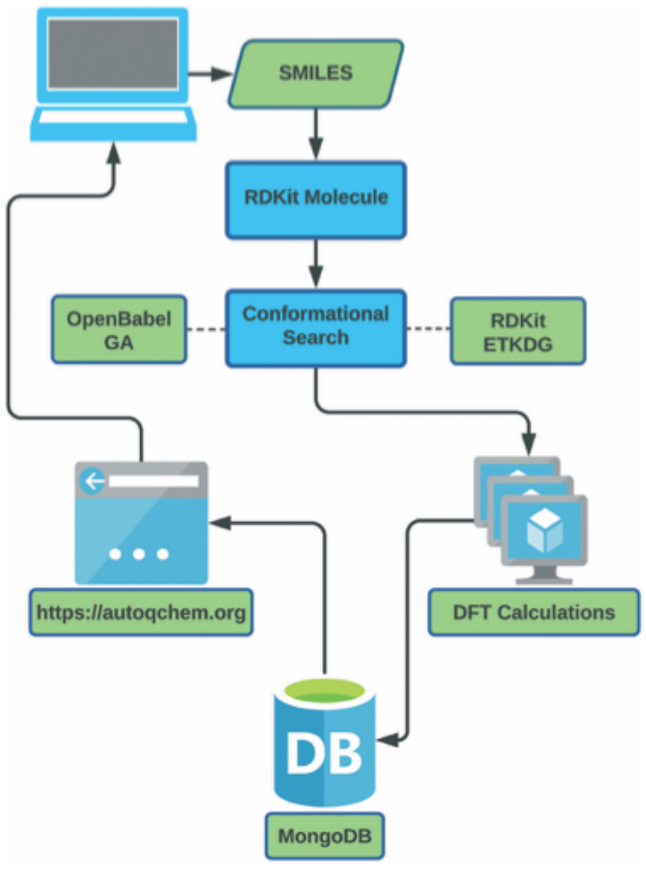


Automated DFT



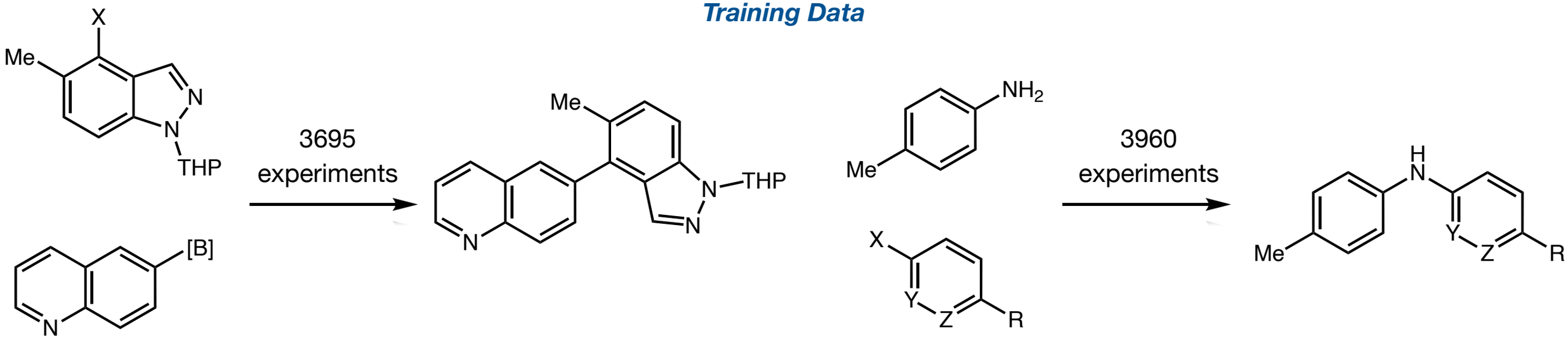
Just need to input SMILES





Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization



For more details about Auto-QChem and other applications see:

Zuranski, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. *React. Chem. Eng.* **2022**, 7, 1276–1284

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

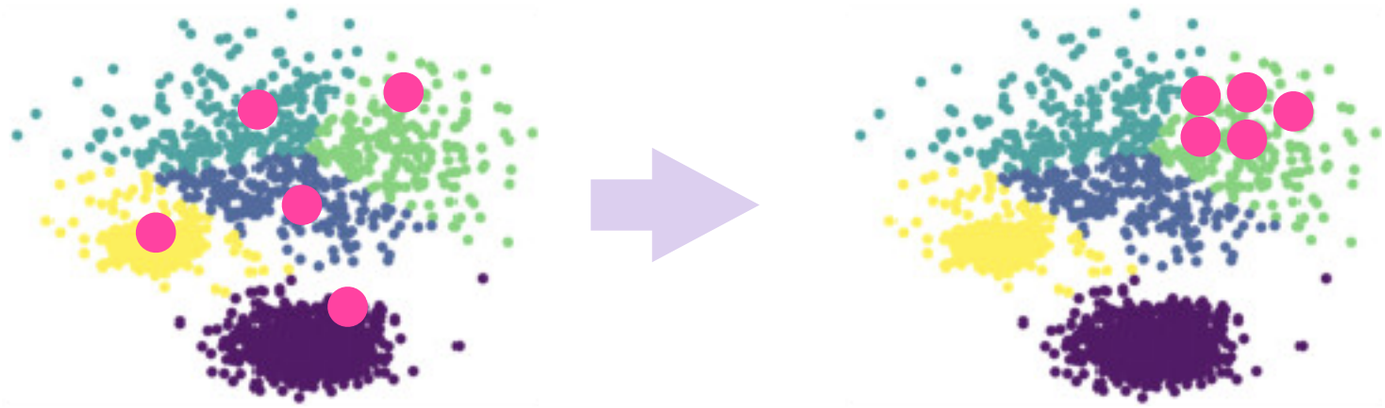
Machine Learning Applied to Chemistry - Reaction Optimization

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Constant mean Gaussian process

Pure Exploiter



Picks experiment expected to give better yield

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

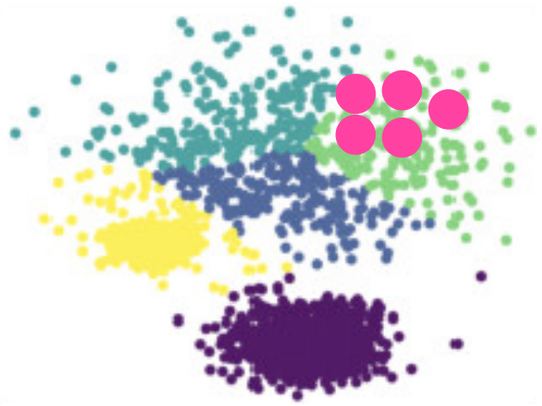
Pioneering acquisition function

Pure Explorer

Picks experiment of greatest predictive uncertainty

Constant mean Gaussian process

Pure Exploiter



Picks experiment expected to give better yield

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

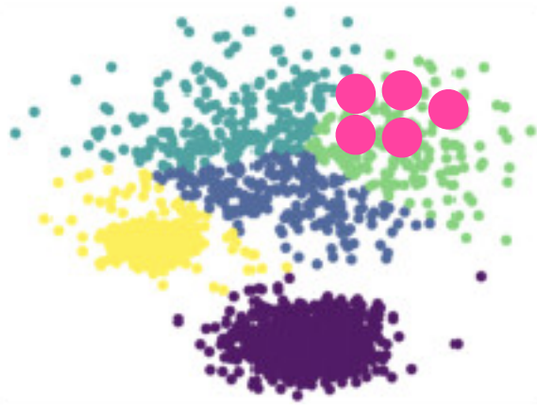
Machine Learning Applied to Chemistry - Reaction Optimization

Pioneering acquisition function

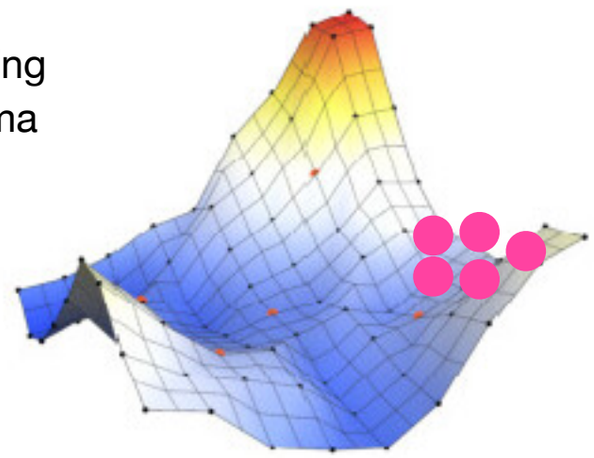
Pure Explorer

Constant mean Gaussian process

Pure Exploiter



Susceptible to becoming trapped in local maxima



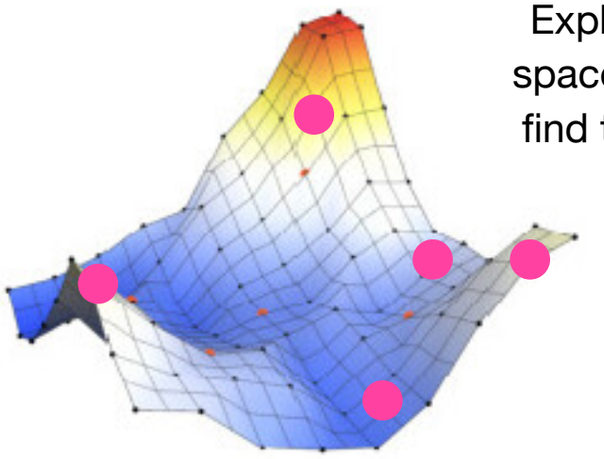
3D representation

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Pioneering acquisition function

Pure Explorer

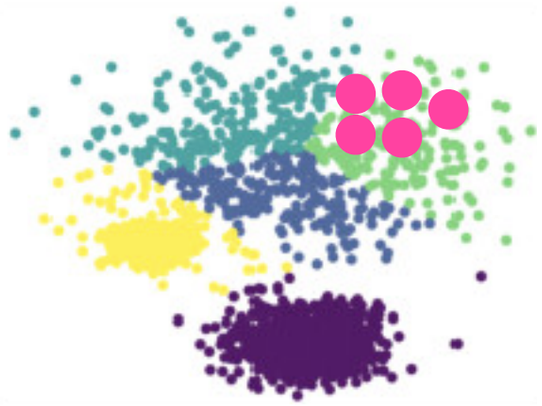


Explores more condition space but not designed to find the absolute maxima

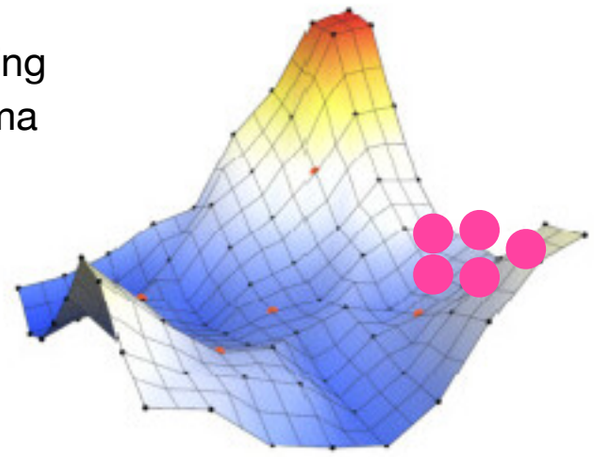
3D representation

Constant mean Gaussian process

Pure Exploiter



Susceptible to becoming trapped in local maxima



3D representation

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

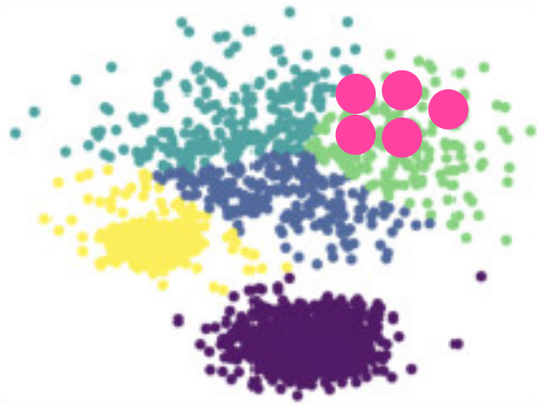
Machine Learning Applied to Chemistry - Reaction Optimization

Pioneering acquisition function

Pure Explorer

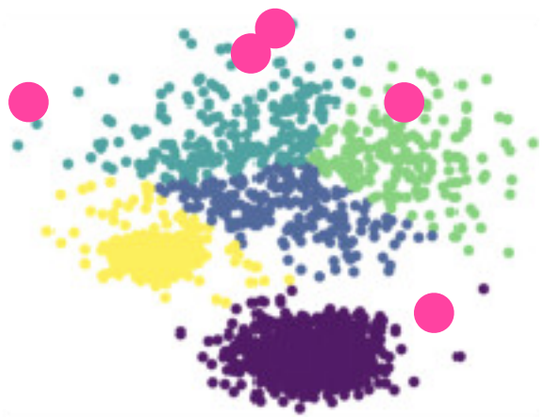
Constant mean Gaussian process

Pure Exploiter



Bayesian optimization with expected improvement

Mix of both



Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

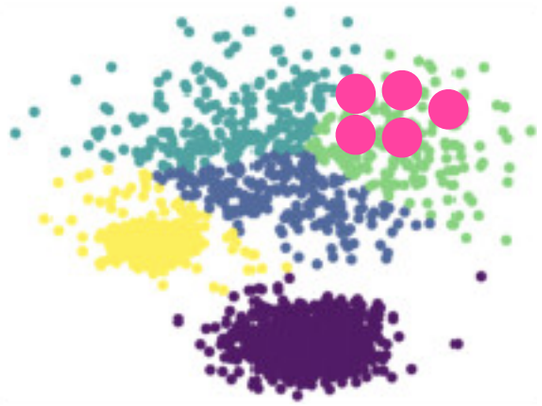
Machine Learning Applied to Chemistry - Reaction Optimization

Pioneering acquisition function

Pure Explorer

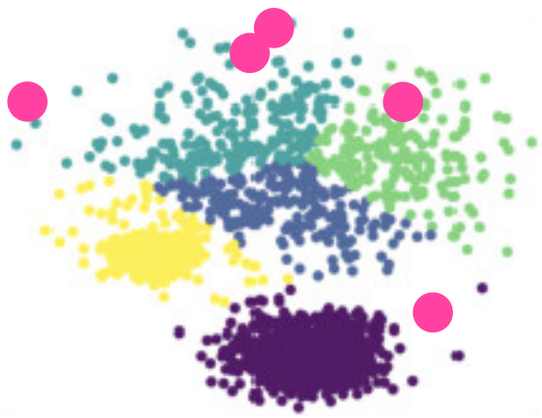
Constant mean Gaussian process

Pure Exploiter

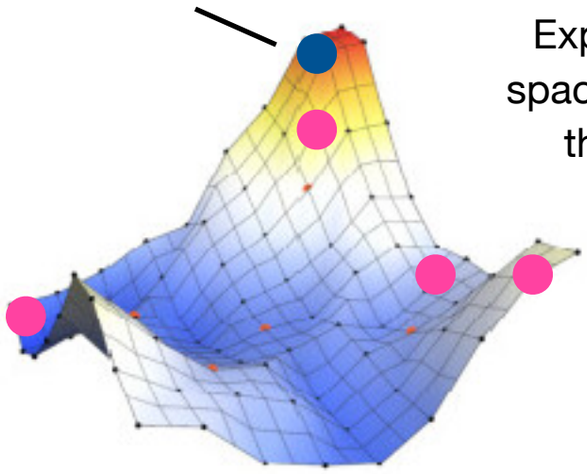


Bayesian optimization with expected improvement

Mix of both



Highest yield

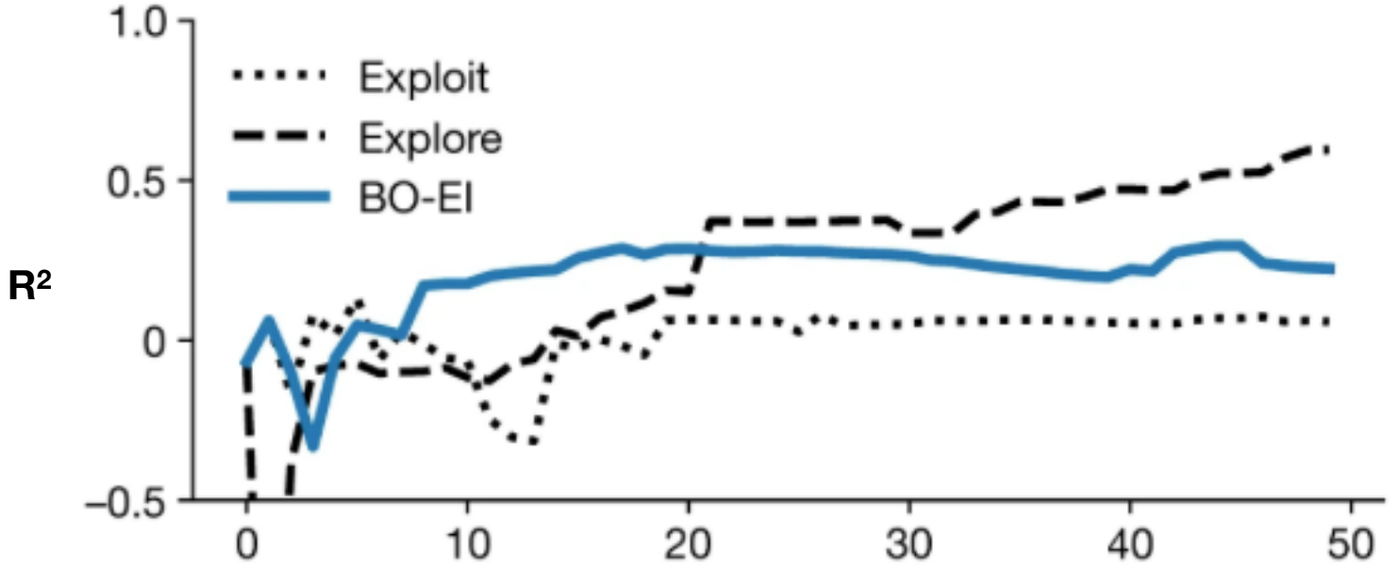


Explores more condition space and designed to find the absolute maxima

3D representation

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

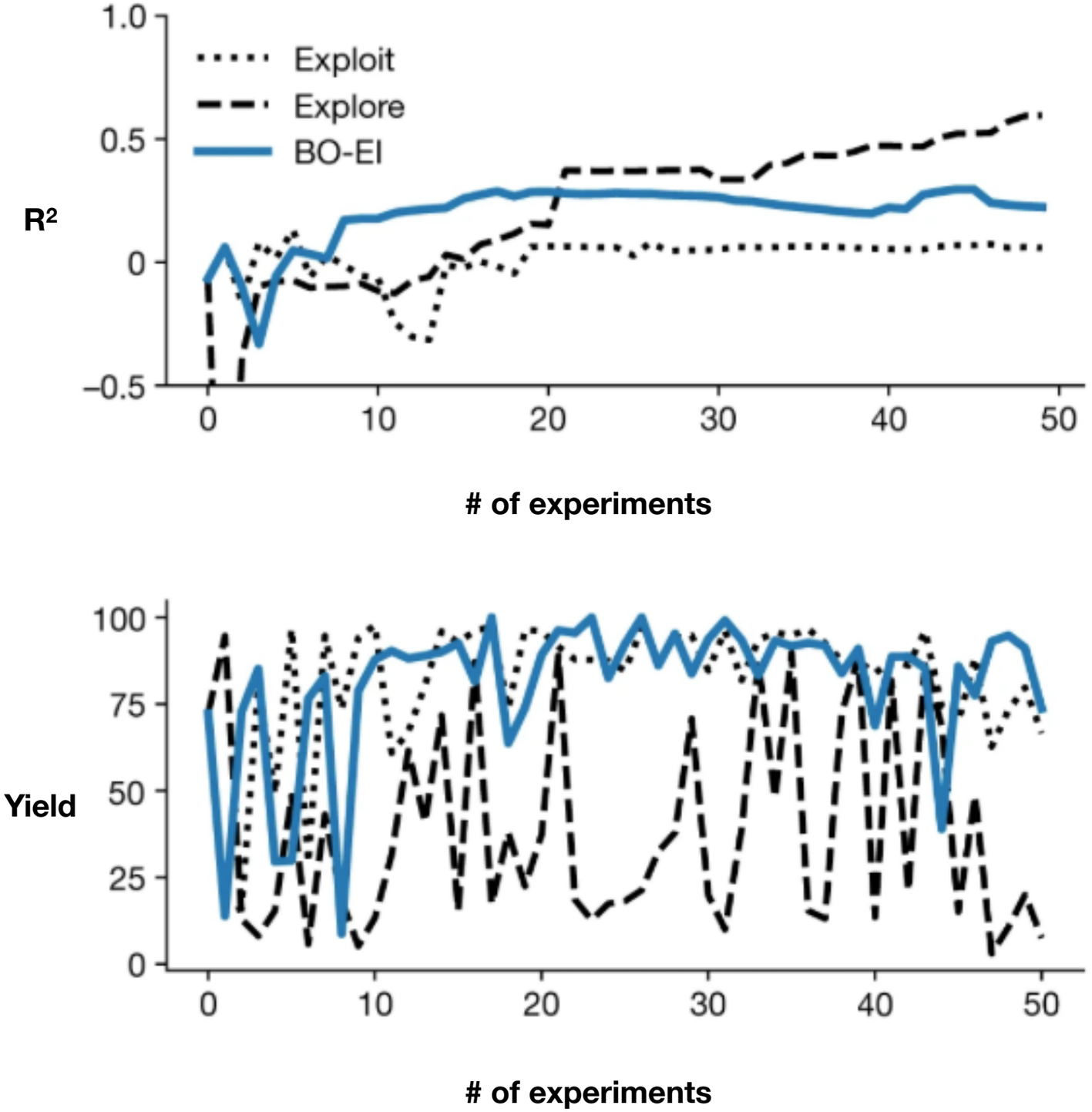
Machine Learning Applied to Chemistry - Reaction Optimization



Explore algorithm and mixed algorithm have a much more accurate "understanding" of the reaction space

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization



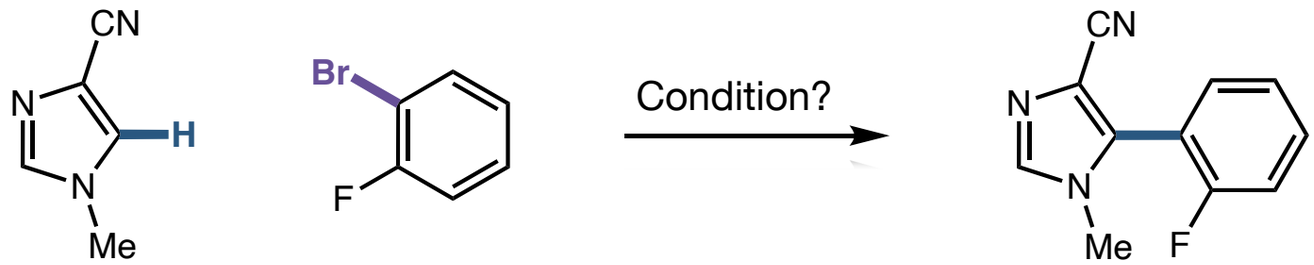
Explore algorithm and mixed algorithm have a much more accurate "understanding" of the reaction space

Explore algorithm doesn't care about highest yield and thus doesn't converge on optimal conditions as well as BO-EI

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Test Case



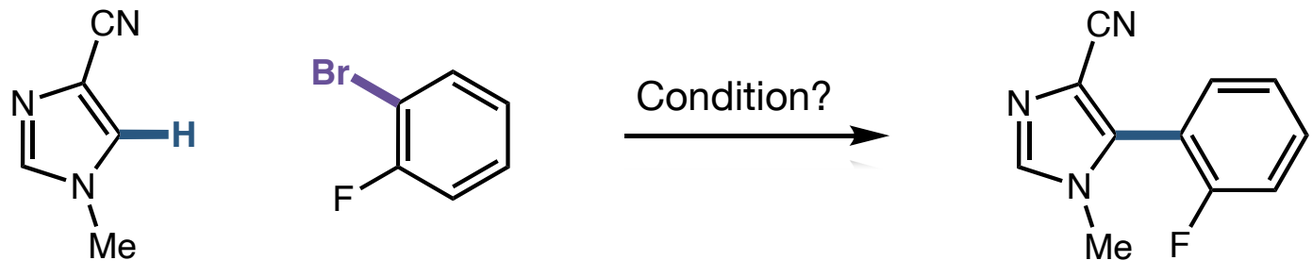
HTE was run for to gather data for the test but ***no data was fed to the model***

Defined a test space of 1728 conditions

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Test Case



*HTE was run for to gather data for the test but **no data was fed to the model***

Defined a test space of 1728 conditions

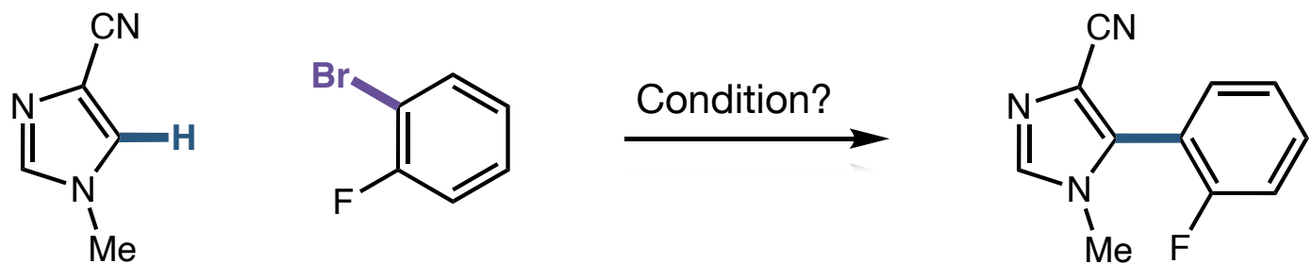
50 expert chemists	**VS**	***50 run of model***

Each player could submit 20 batches of 5 experiments, and between each batch get the results from the HTE data

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

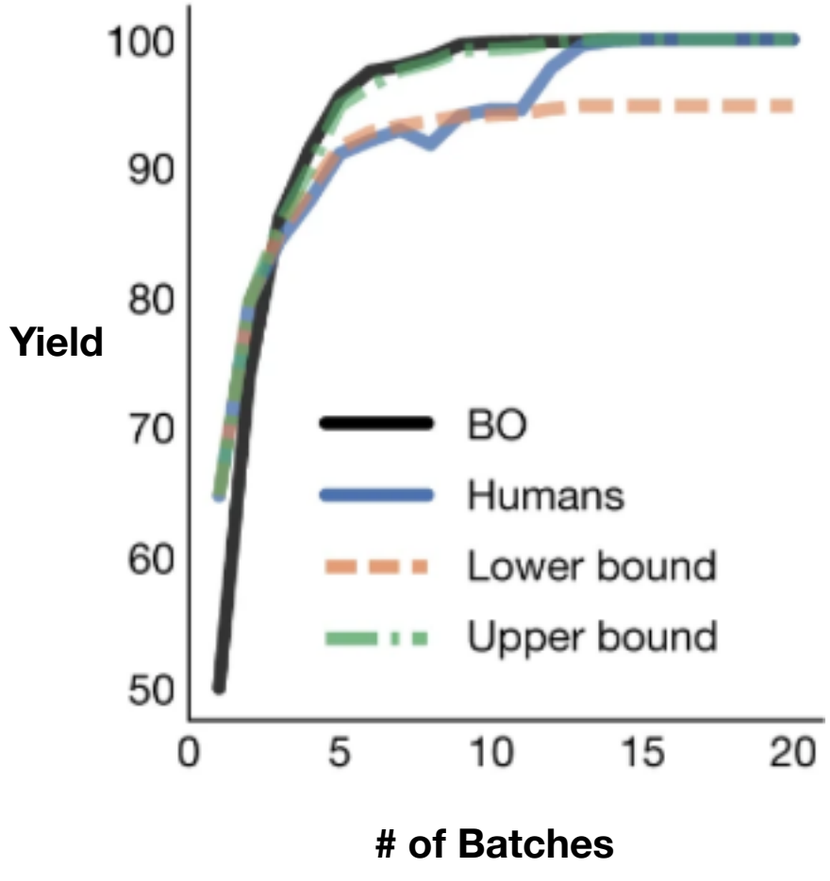
Test Case



*HTE was run for to gather data for the test but **no data was fed to the model***

Defined a test space of 1728 conditions

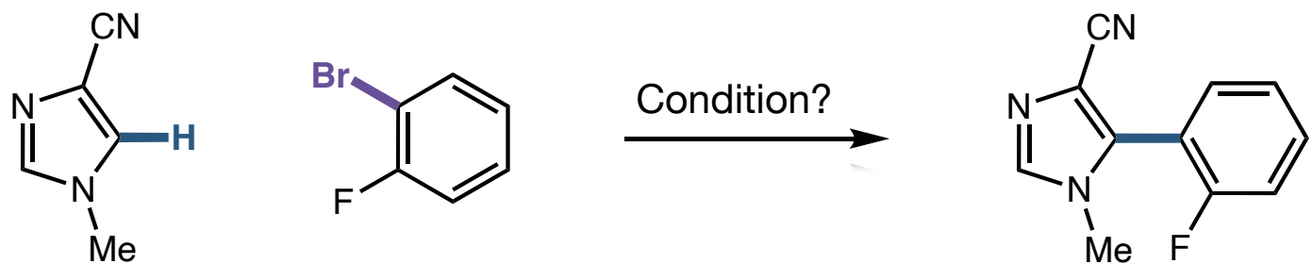
50 expert chemists **VS** ***50 run of model***



Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

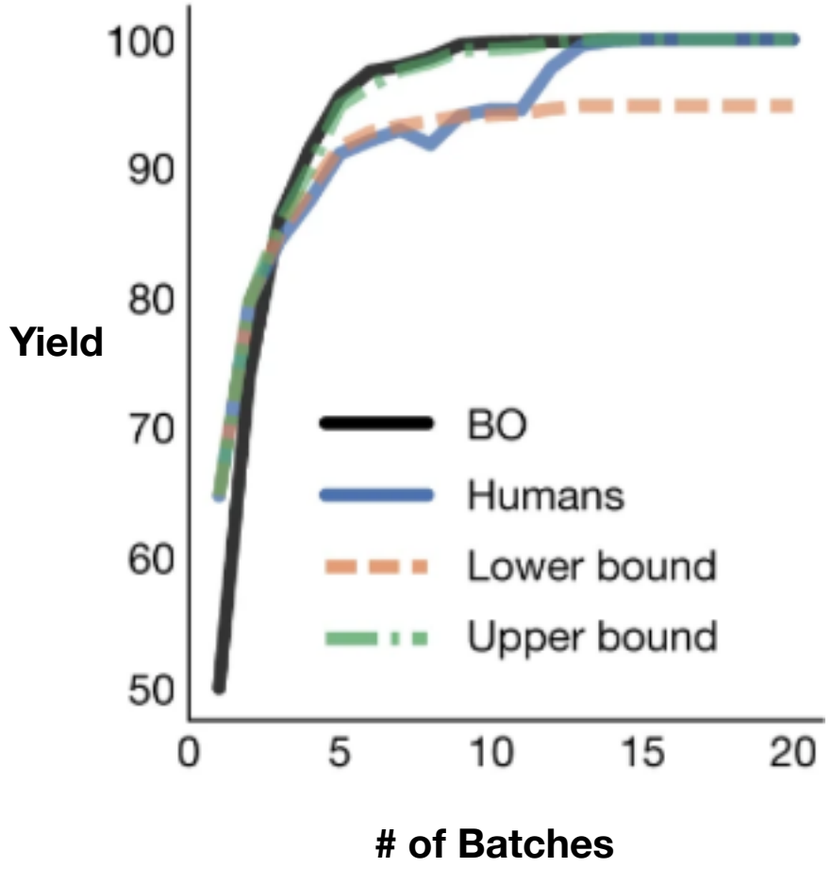
Test Case



*HTE was run for to gather data for the test but **no data was fed to the model***

Defined a test space of 1728 conditions

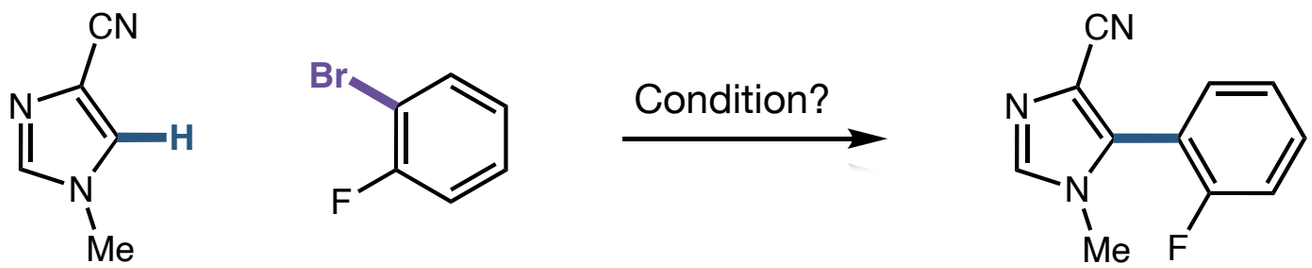
On average the model outperformed experts by the 5th batch of experiments



Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Test Case



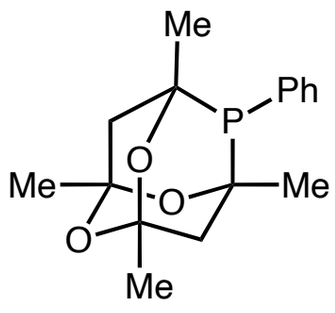
HTE was run for to gather data for the test but ***no data was fed to the model***

Defined a test space of 1728 conditions

Unprecedented ligand for this class of reaction

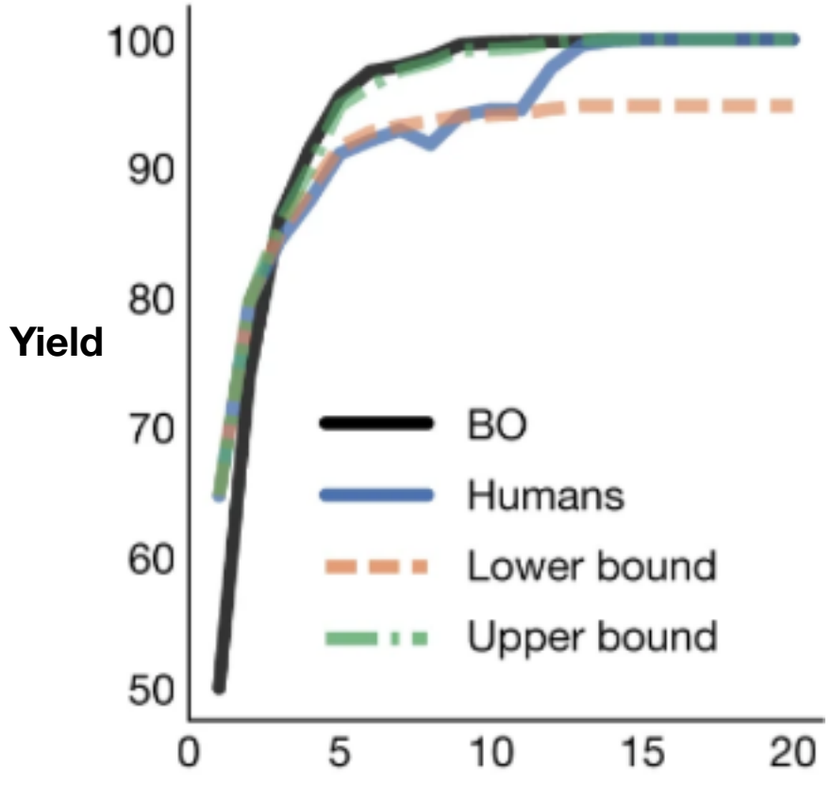
Difficult for the experts to "find" this ligand

Optimal Conditions



$PdCl_2(allyl)_2$
CsOPiv, 105 °C
DMA (0.153 M),

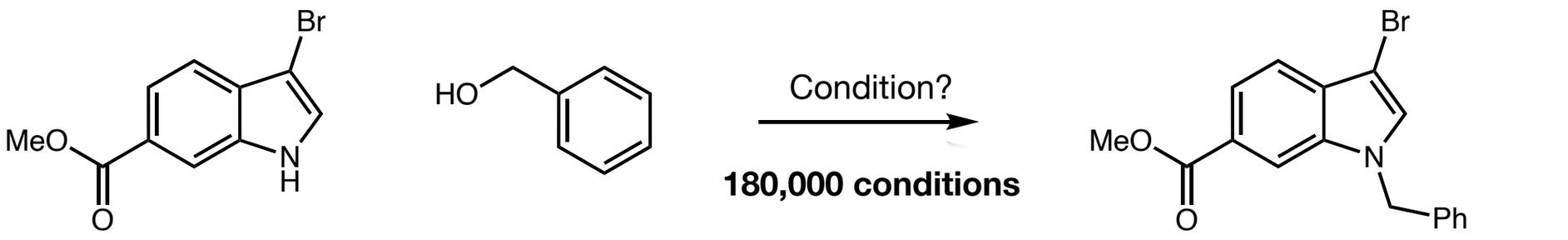
100% yield
Less than 50 experiments for all 50 different starting points



Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

Application to Mitsunobu reaction



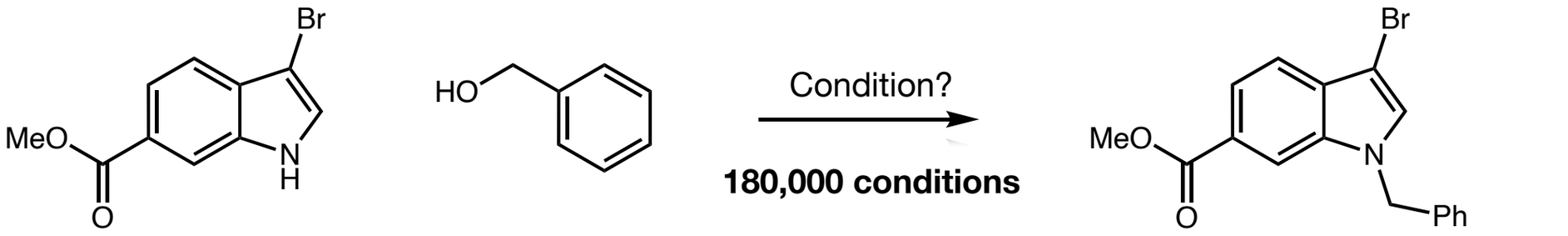
30 experiments
99% yield

Standard Conditions: 60% yield

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

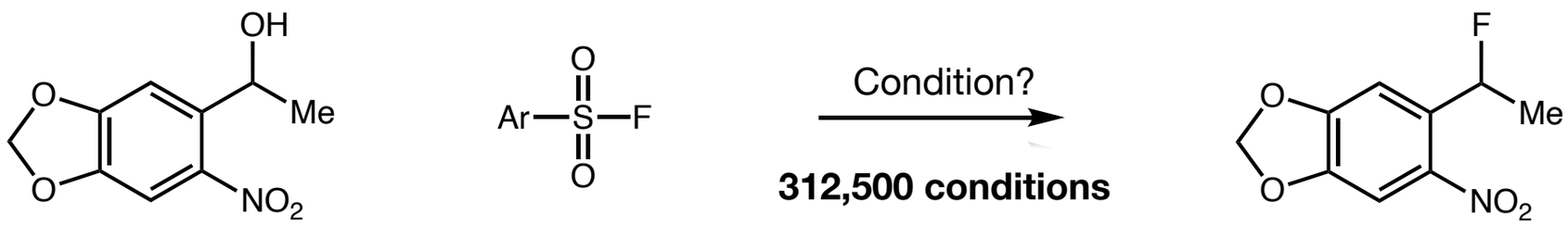
Application to Mitsunobu reaction



30 experiments
99% yield

Standard Conditions: 60% yield

Application to deoxyfluorination reaction

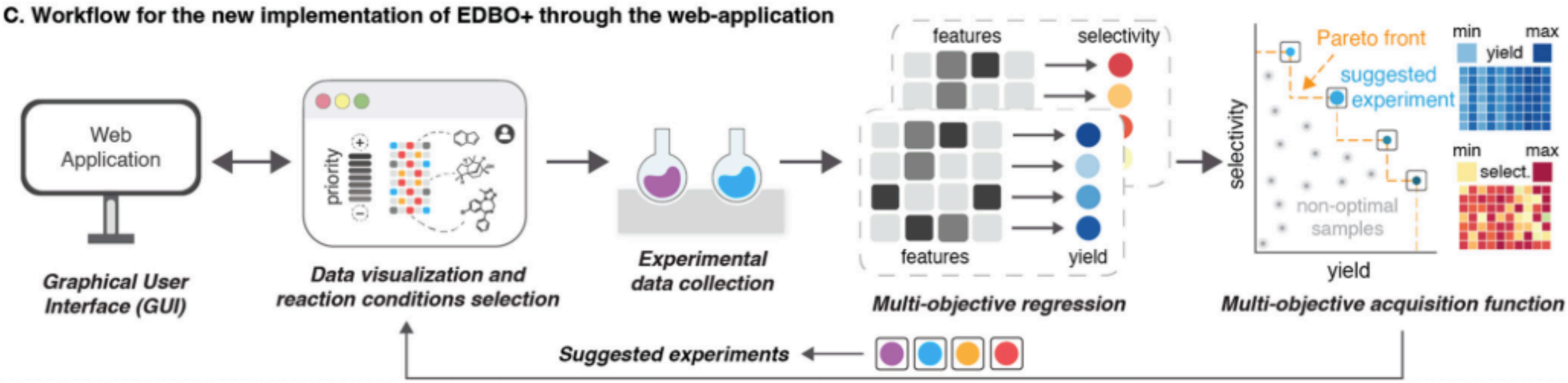


40 experiments
70% yield

Standard Conditions: 36% yield

Shields, B. J.; Stevens, J.; Li, Jun.; Parasram, M.; Damani, F.; Alvarado, Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89–96.

Machine Learning Applied to Chemistry - Reaction Optimization

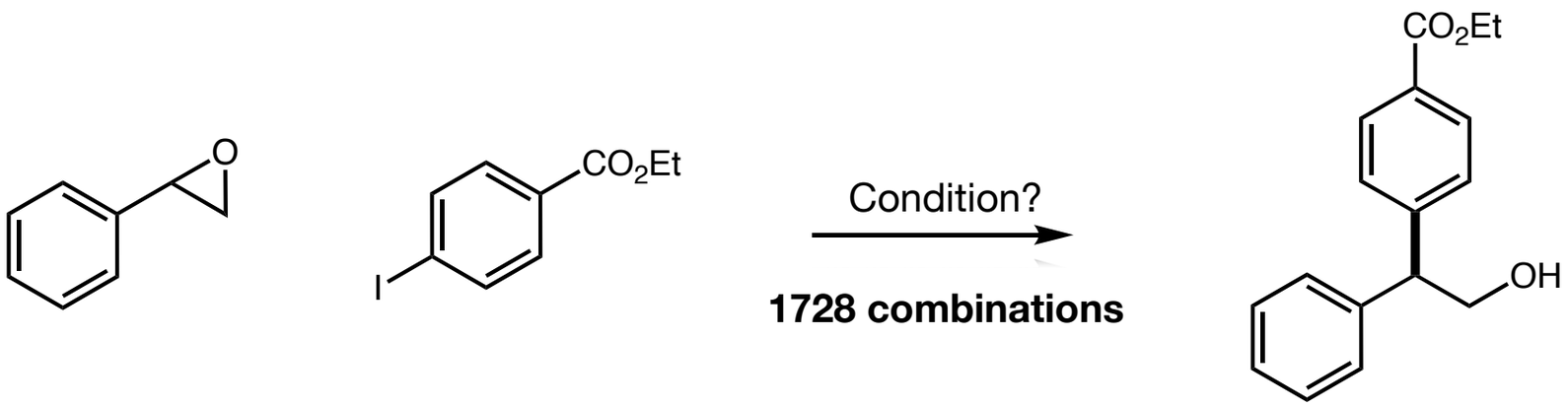


Website to use EDBO:
https://www.edbowebapp.com

Doyle, A. G. *J. Am. Chem. Soc.* **2022**, 144, 19999–20007.

Machine Learning Applied to Chemistry - Reaction Optimization

Bayesian Optimization applied to enantioselective photocatalytic systems

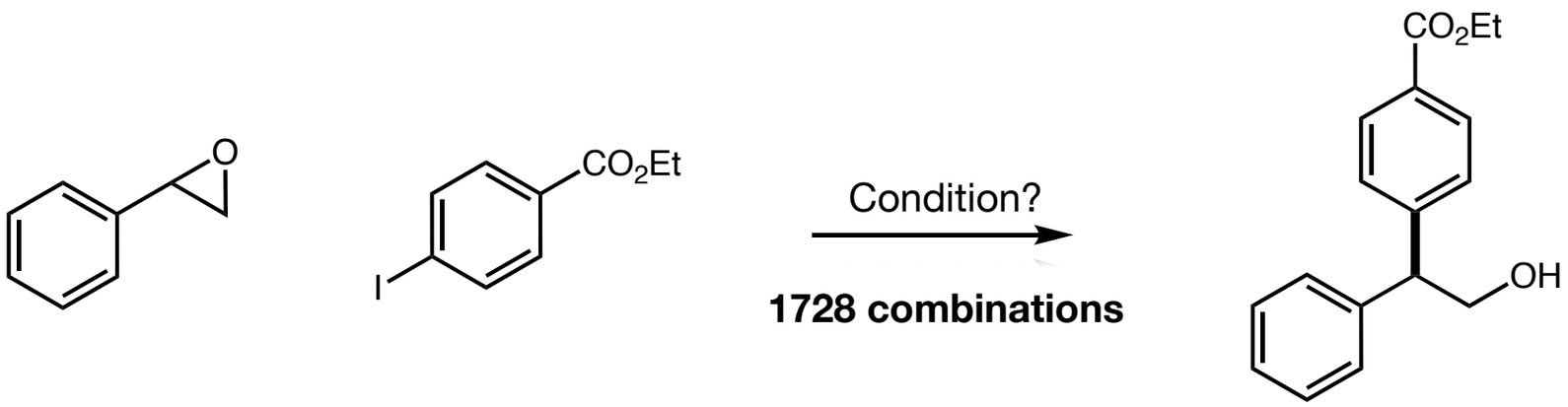


Human optimization: ~500 reactions for 70% yield, 80% ee

Doyle, A. G. *J. Am. Chem. Soc.* **2022**, 144, 19999–20007.

Machine Learning Applied to Chemistry - Reaction Optimization

Bayesian Optimization applied to enantioselective photocatalytic systems



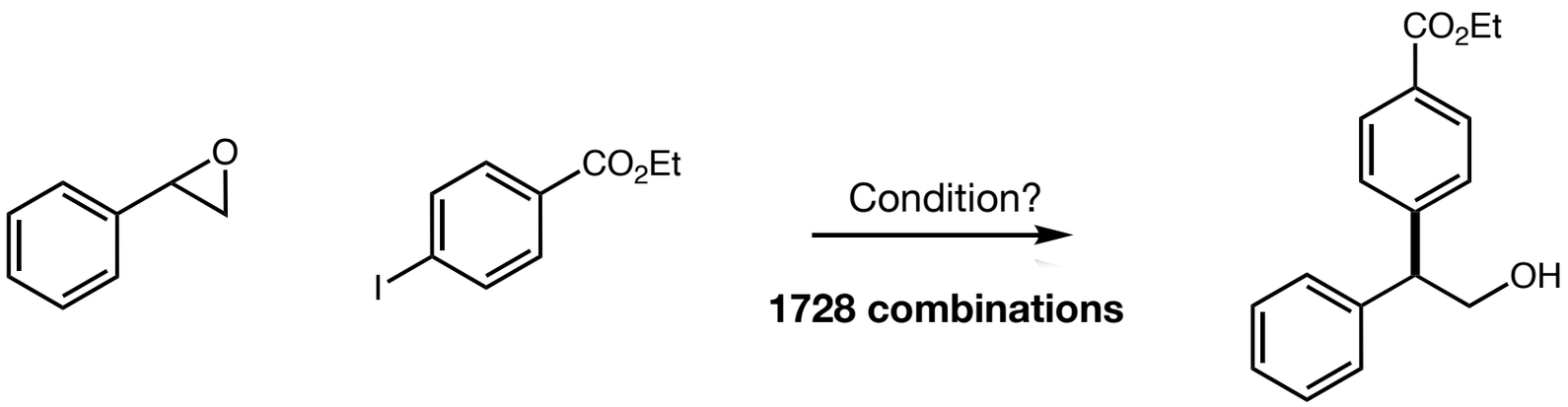
24 experiments
80% yield, 91% ee

Human optimization: ~500 reactions for 70% yield, 80% ee

Doyle, A. G. *J. Am. Chem. Soc.* **2022**, 144, 19999–20007.

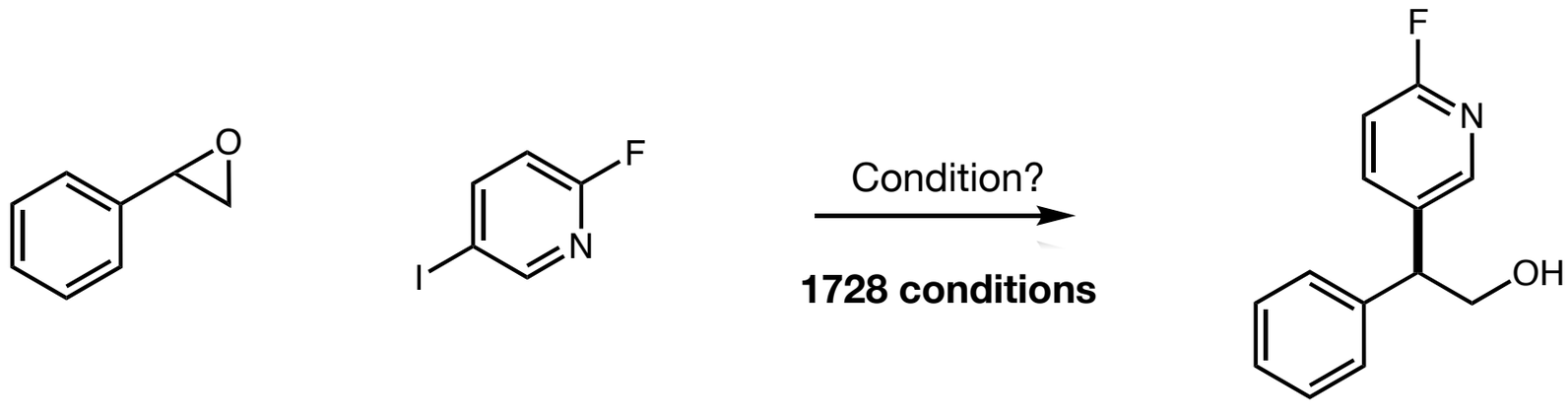
Machine Learning Applied to Chemistry - Reaction Optimization

Bayesian Optimization applied to enantioselective photocatalytic systems



24 experiments
80% yield, 91% ee

Human optimization: ~500 reactions for 70% yield, 80% ee

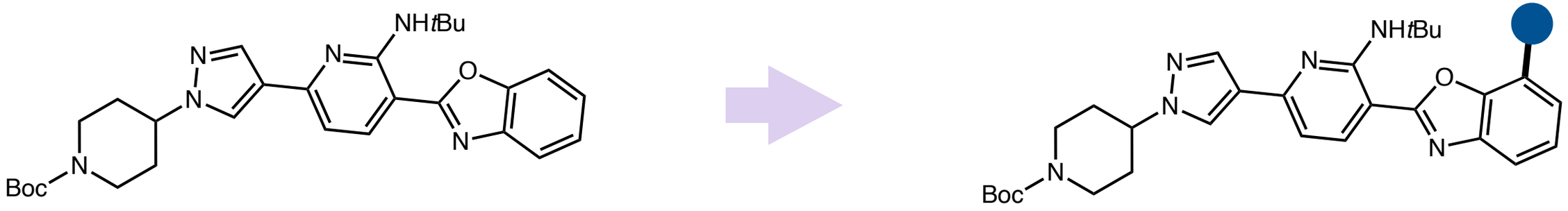


15 experiments
59% yield, 77% ee

Human optimization: 49% yield, 76% ee

Doyle, A. G. *J. Am. Chem. Soc.* **2022**, 144, 19999–20007.

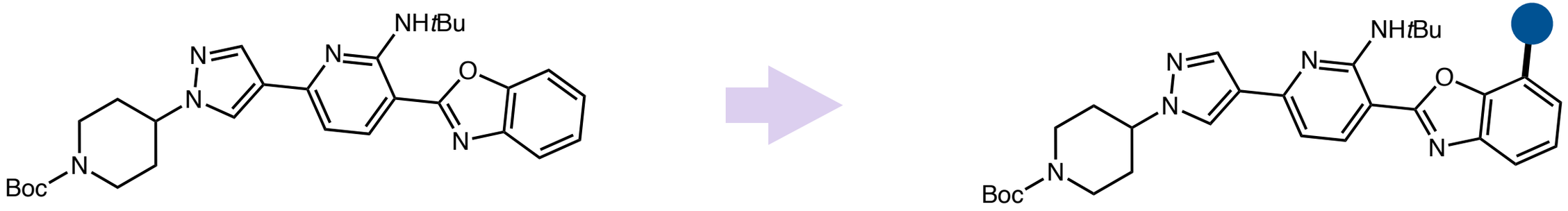
Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings



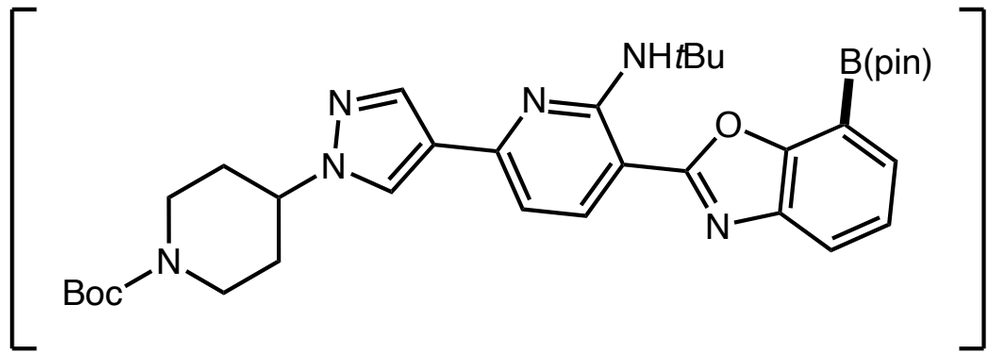
Theoretical Question: How can we access substitution at this position to fill a binding pocket?

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings



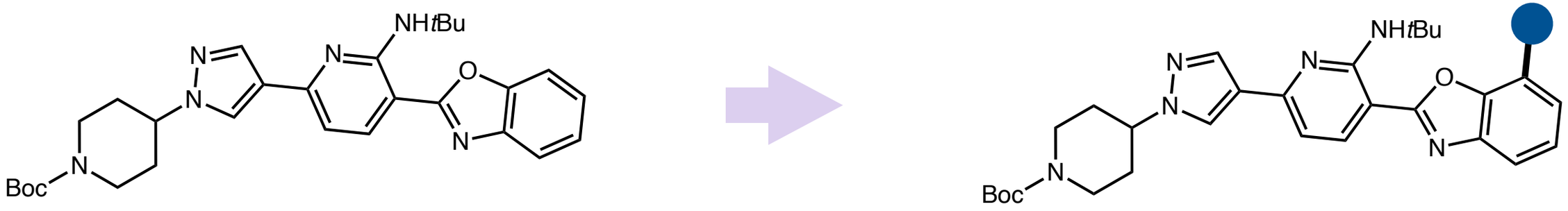
Theoretical Question: How can we access substitution at this position to fill a binding pocket?



Late-stage functionalization to lynchpin intermediate

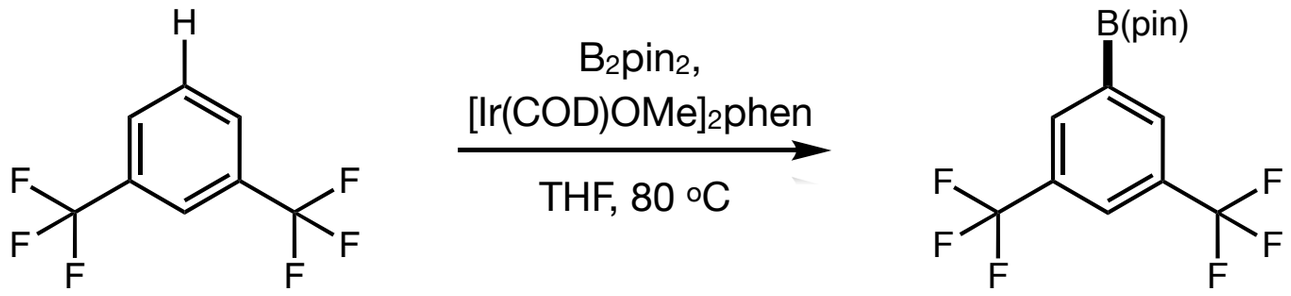
Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings



Theoretical Question: How can we access substitution at this position to fill a binding pocket?

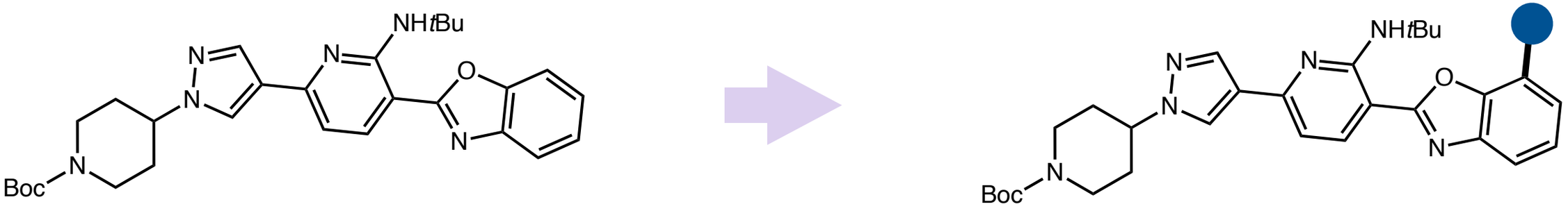
Hartwig C-H Borylation



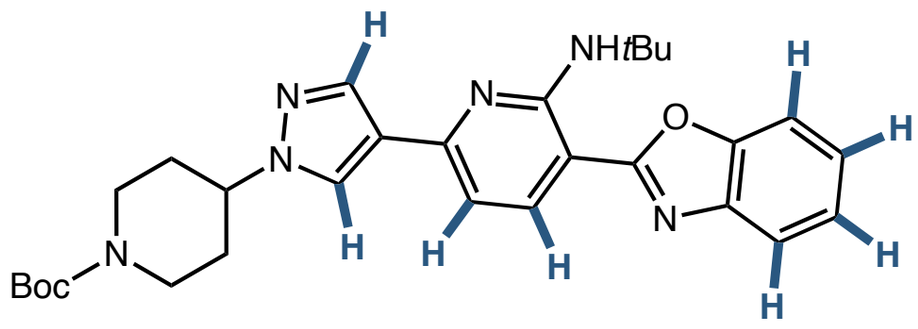
For simple systems selectivity can be a priori predicted by empirical knowledge of selectivity from other substrates

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings



Theoretical Question: How can we access substitution at this position to fill a binding pocket?

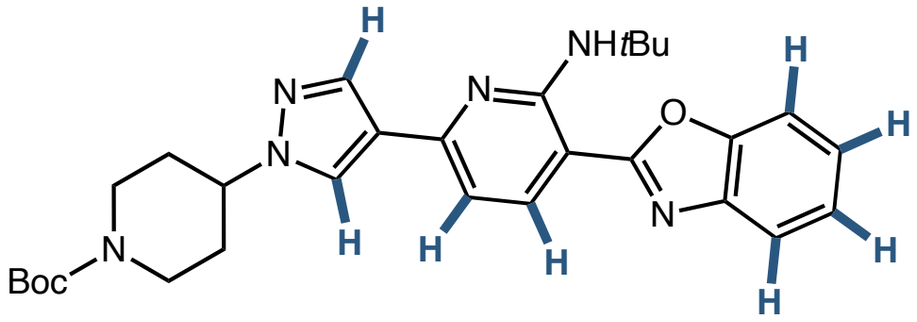


Many C-H bonds: predicting selectivity is not trivial

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

Task

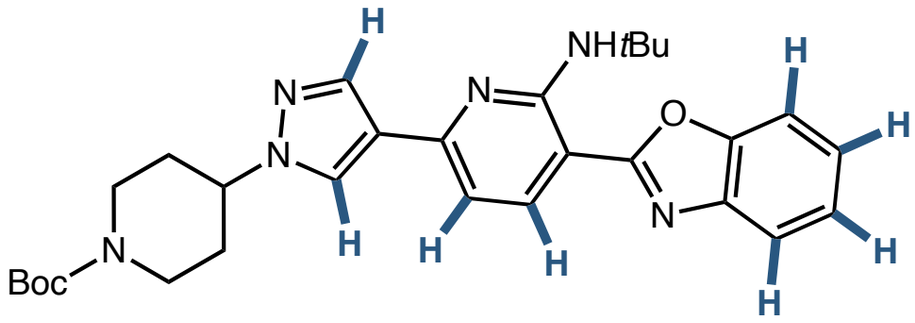


What C-H bond will be borylated?

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

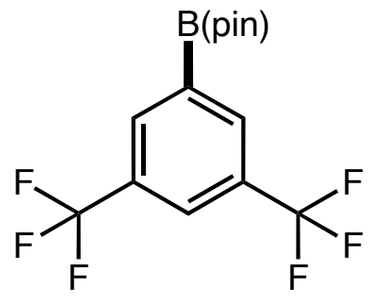
Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

Task



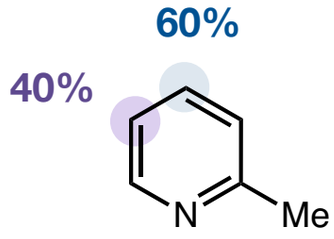
What C-H bond will be borylated?

Data Set



86 literature examples

+

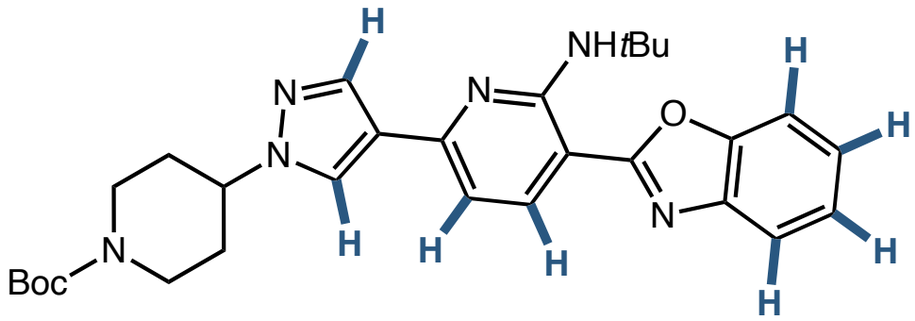


15 less-selective examples

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

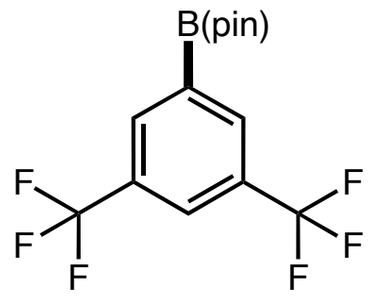
Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

Task



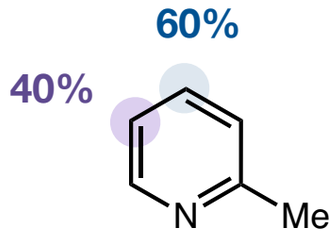
What C-H bond will be borylated?

Data Set



86 literature examples

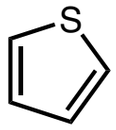
+



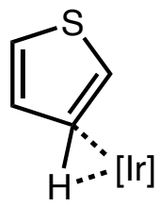
15 less-selective examples

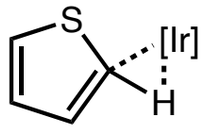
Data parameterization

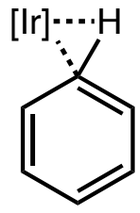
Input SMILES



DFT TS for each C-H bond





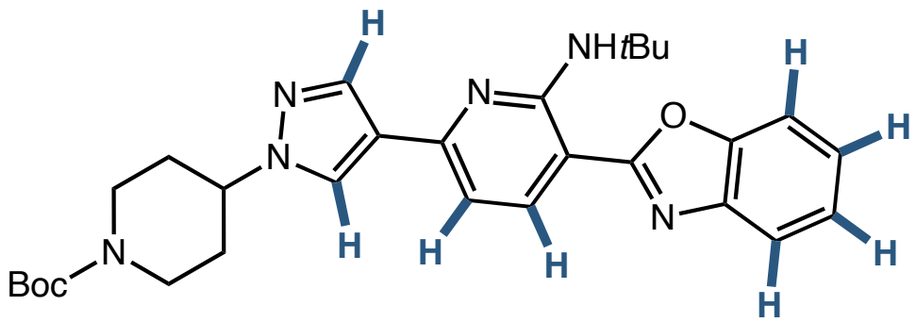


Ph used as reference

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

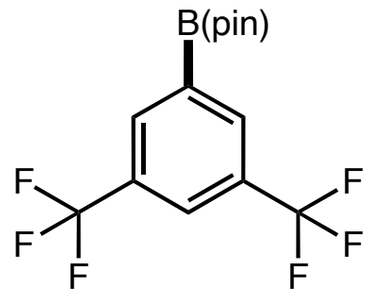
Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

Task



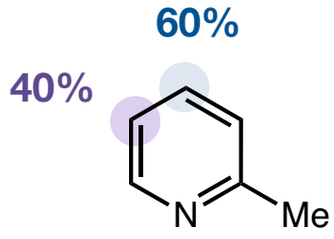
What C-H bond will be borylated?

Data Set



86 literature examples

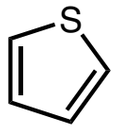
+



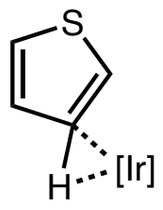
15 less-selective examples

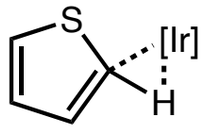
Data parameterization

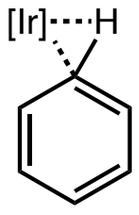
Input SMILES



DFT TS for each C-H bond







Ph used as reference

Modeling

Partial Least Squares algorithm (electronic)

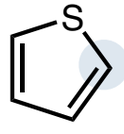
+

Sterimol Parameters (steric)

=

Predicted relative energy barrier

Boltzmann weights give predicted product distribution

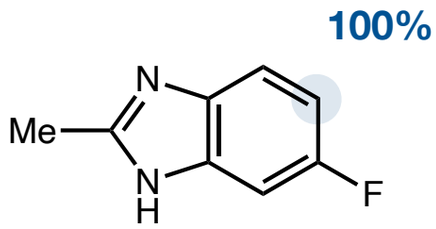


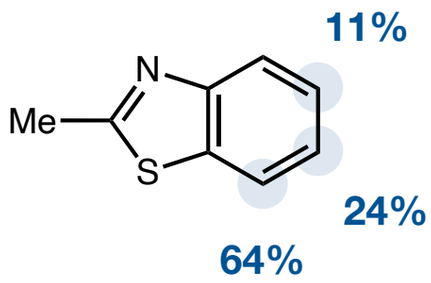
100%

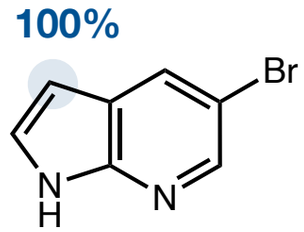
Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

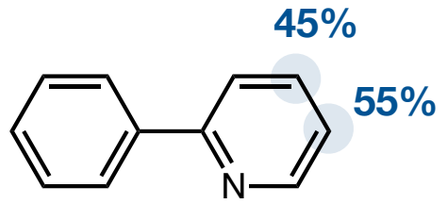
Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

Experimental Selectivities

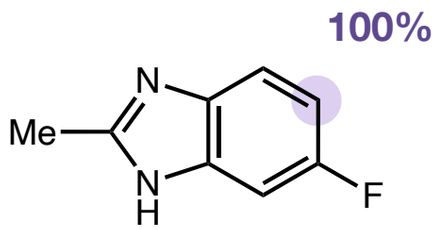


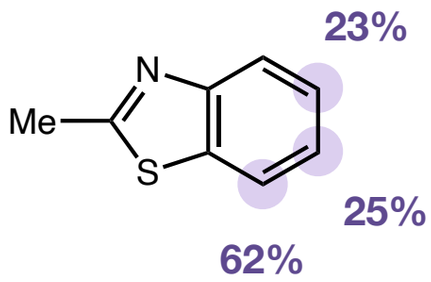


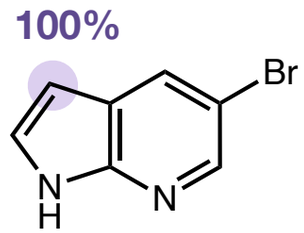


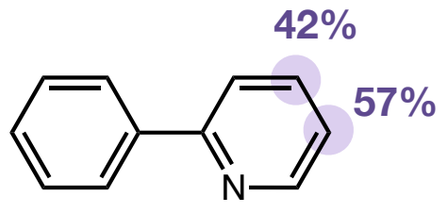


Predicted Selectivities









Model is quite accurate in predicting selectivity for simple substrates

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

predict the major site of borylation

15 expert chemists

vs

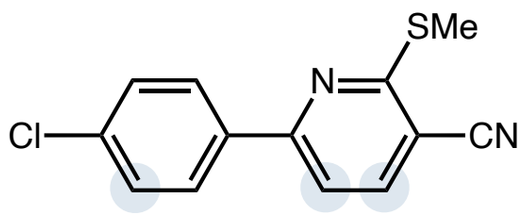
Model

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

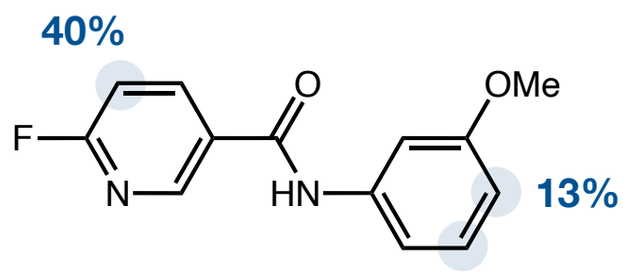
Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

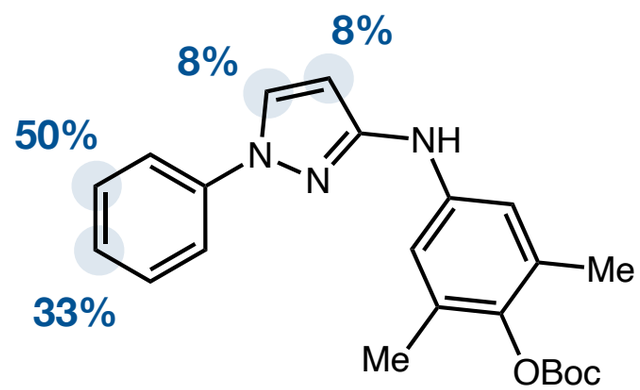
predict the major site of borylation

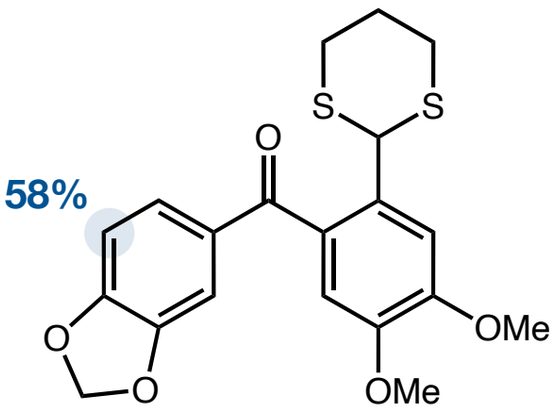
15 expert chemists **vs** ***Model***



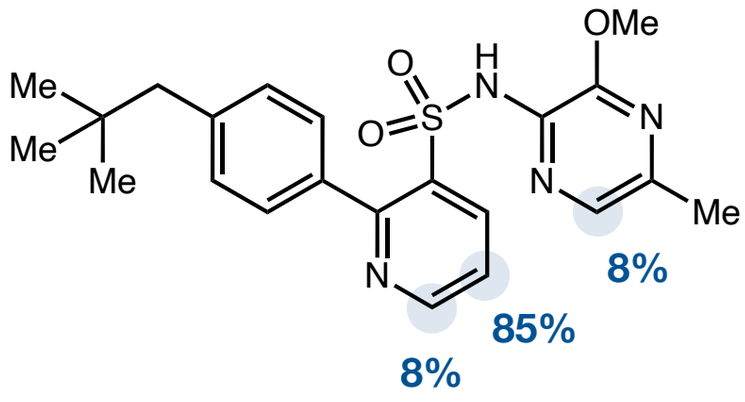
No reaction: 46%

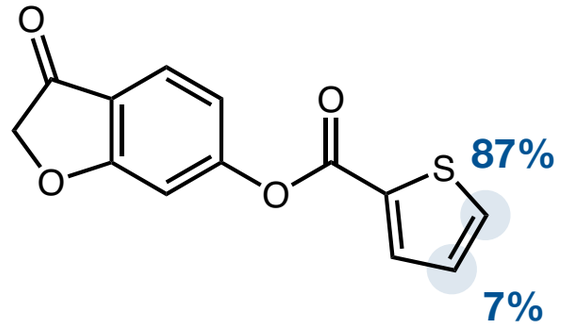






No reaction: 42%





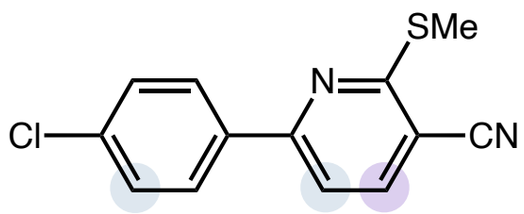
No reaction: 7%

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

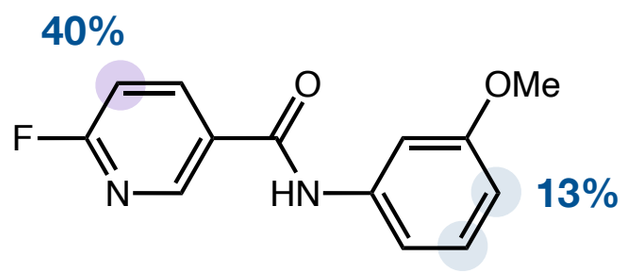
Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

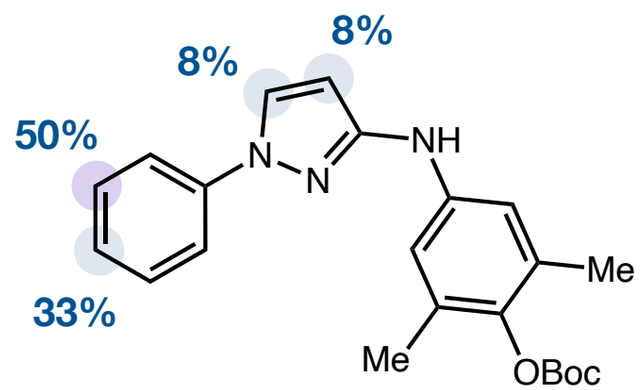
predict the major site of borylation

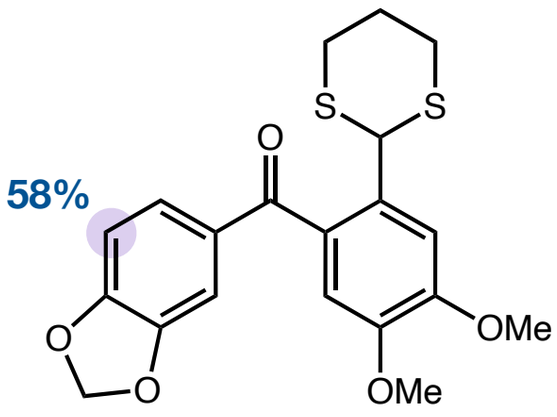
15 expert chemists **vs** ***Model***



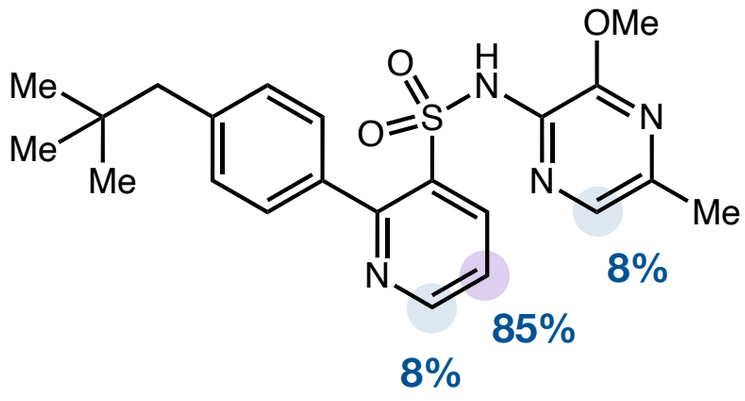
No reaction: 46%

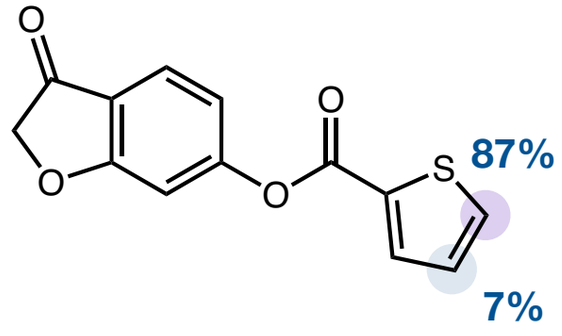






No reaction: 42%





No reaction: 7%

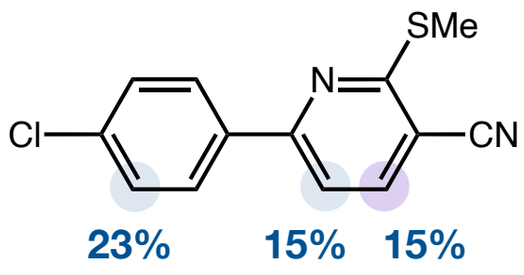
Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

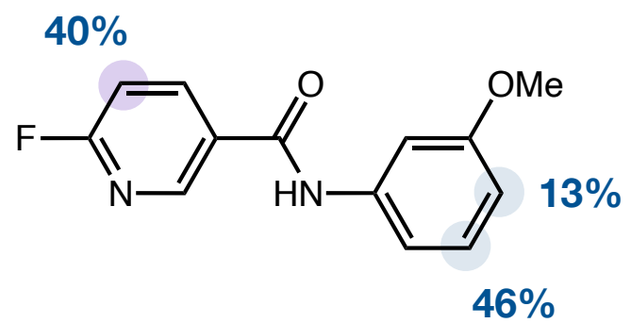
predict the major site of borylation

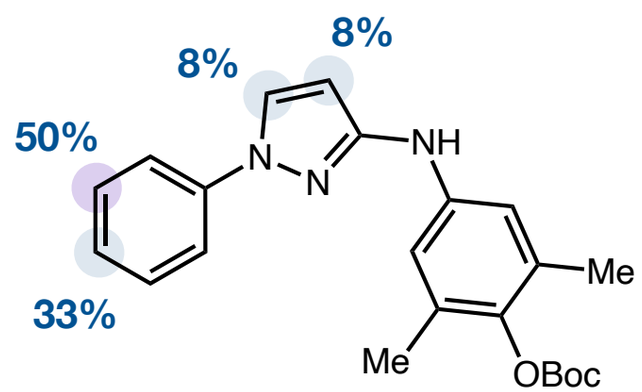
15 expert chemists **vs** ***Model***

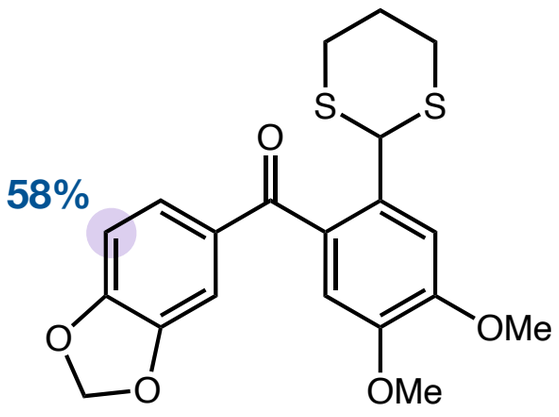
Correct answer in purple



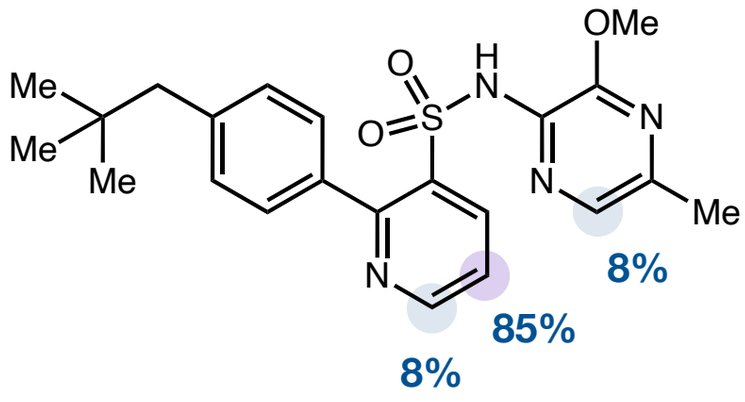
No reaction: 46%

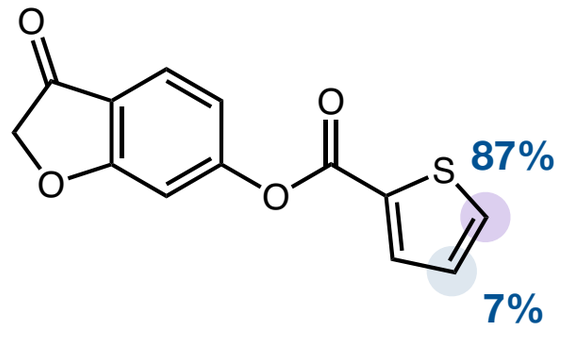






No reaction: 42%





No reaction: 7%

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning Applied to Chemistry - Selectivity Predictions in Complex Settings

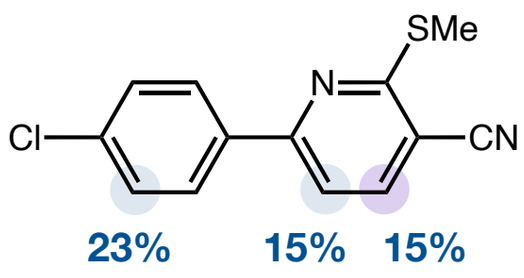
Overall: 56% accurate

15 expert chemists

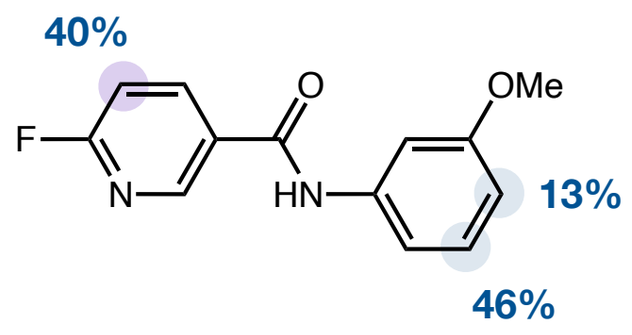
VS

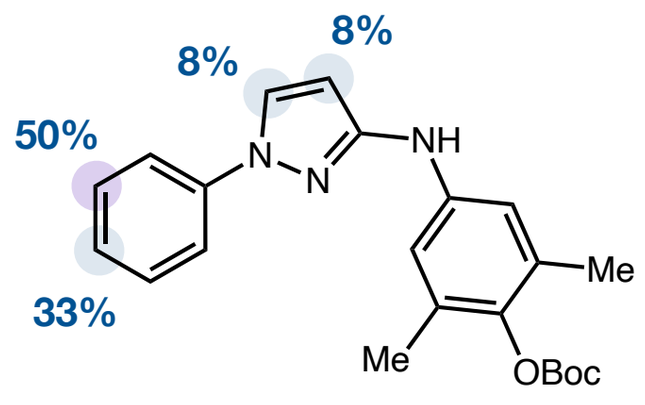
Model

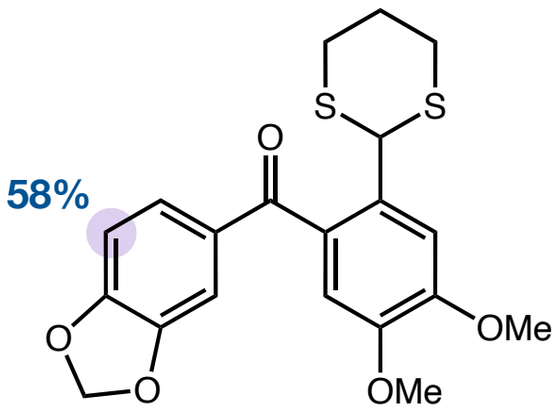
Overall: 100% accurate



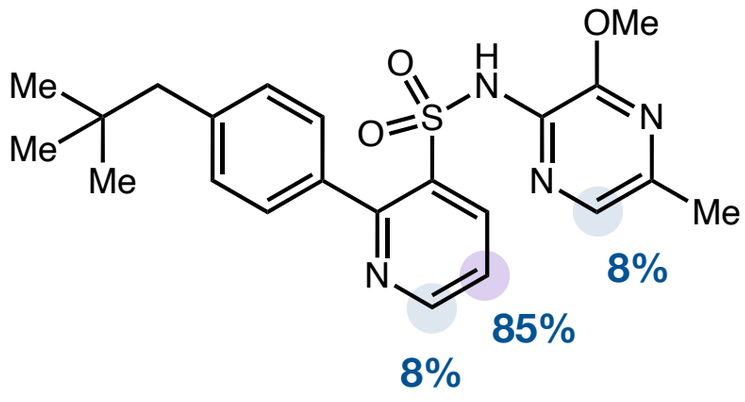
No reaction: 46%

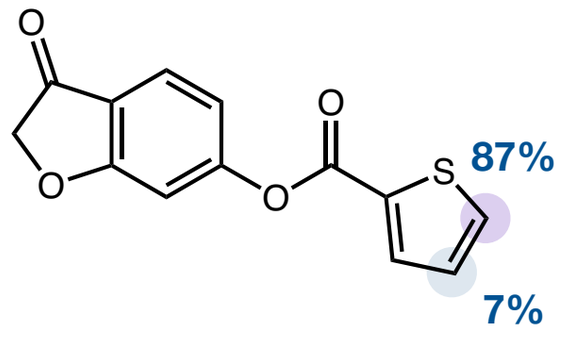






No reaction: 42%





No reaction: 7%

Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Skold, C.; Norrby, P.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023**, *145*, 17367–17376.

Machine Learning in Chemistry

Reviews

Zuranski, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. *Acc. Chem. Res.* **2021**, *54*, 1856–1865

Meuwly, M. *Chem. Rev.* **2021**, *121*, 10218–10239

Tu, Z.; Stuyver, T.; Coley, C. W. *Chem Sci*, **2022**, *14,* 226–244

Aal E Ali, R. S.; Meng, J.; Khan, M. E. I.; Jiang, X. *Artificial Intelligence Chemistry*, **2024**, *2*

Additional Examples

Romer, N. P,.; Min, D. S.; Wang, J. Y.; Walroth, R. C.; Mack, K. A.; Sirois, L. E.; Gosselin, F.; Zell, D.; Doyle, A. G.; Sigman, M. S. *ACS Catal.* **2024**, *14*, 4699–4708

Wang, J. Y. […] Doyle, A. G. *Nature*, **2024**, *626*, 1025–1033

Zuranski, A. M.; Gandhim S. S.; Doyle, A. G. *J. Am. Chem. Soc.* **2023**, *145*, 7898–7909

Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science*, **2019**, *247*

Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A .S.; Bigler, R.; Fantasia, S.; Denmark, S. E. **2023**, *381*, 965–972

Baczewska, P.; Kulczykowski, M.; Zambron, B.; Adamczak, J.; Pakulski, Z.; Roszak, R.; Grzybowski, B. A.; Młynarski, J. *Angew. Chen. Int. Ed.* **2024**, *136*

Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. *Acs Cent. Sci.* **2018**, *4*, 1465–1476

Angello, N. H.; Rathore, V.; Beker, W.; Wolos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroder, C. M.; Guzik, A .A.; Grzybowski, B. A.; Burkę, M. D. *Science,* **2022**, *378*, 399–405

Limitations

Schnitzer, T.; Schnurr, M.; Zahrt, A. F.; Sakhaee, N.; Denmark, S. E.; Wennemers, H. *Acs Cent. Sci*, **2024**, *10*, 367–373

Beker, W.; Roszak, R.; Wolos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grybowski, B. A. *J. Am. Chem. Soc.* **2022**, *144*, 4819–1827

The path to developing a useful method

1. Identify a problem

Accelerated serendipity

2. Initial hit

Accelerated serendipity

Modern HTE

3. Optimization

Bayesian optimization

ML based virtual screening

Multidimensional LFER

4. Generalize

ML prediction of optimal catalyst

Additive screening strategies

General HTE

Data science substrate mapping

5. General adoption

ML prediction of selectivity

ML prediction of yield

The path to developing a useful method

1. Identify a problem

Accelerated serendipity

2. Initial hit

Accelerated serendipity

Modern HTE

3. Optimization

Bayesian optimization

ML based virtual screening

Multidimensional LFER

Questions?

4. Generalize

ML prediction of optimal catalyst

Additive screening strategies

General HTE

Data science substrate mapping

5. General adoption

ML prediction of selectivity

ML prediction of yield