www.sciencemag.org/cgi/content/full/science.aar5169/DC1

Supplementary Material for

Predicting reaction performance in C–N cross-coupling using machine learning

Derek T. Ahneman,[1] Jesús G. Estrada,[1] Shishi Lin,[2] Spencer D. Dreher,[2]*
Abigail G. Doyle[1]*

*Corresponding author. Email: spencer_dreher@merck.com (S.D.D.); agdoyle@princeton.edu (A.G.D.)

Published 15 February 2018 as *Science* First Release
DOI: 10.1126/science.aar5169

This PDF file includes:

Materials and Methods
Supplementary Text
Figures S1 to S34
Tables S1 to S4
References

Predicting reaction performance in C–N cross-coupling using machine learning

Derek T. Ahneman,[1] Jesús G. Estrada,[1] Shishi Lin,[2] Spencer D. Dreher,[2]* and Abigail G. Doyle[1]*

*Corresponding authors. Email: spencer_dreher@merck.com (S.D.D.) and agdoyle@princeton.edu (A.G.D.)

Supplementary Materials

Contents

I. General Information S3
II. Ultra High Throughput Screening Data S5
Reaction Setup **S5**
Plate Layout **S6**
Sample Yield Determination **S9**
Plate Heatmaps **S11**
III. Programming Details S13
Instructions for Using `rxnpredict` **S13**
How the Program Works **S15**
Modeling **S25**
IV. Isoxazole Preparation S42
V. Products S44
VI. Organometallic Results S46
VII. NMR Spectra S51

I. General Information

Materials. Reagents were purchased from commercial suppliers (Aldrich, Fisher, Enamine, Combi-Blocks) and used with no further purification. The ligands and bases were stored in a glovebox. Solvents were purchased from Aldrich, anhydrous, sure-seal quality, and used with no further purification. All reactions were set up inside an MBraun glovebox operating with a constant N_2 purge (oxygen typically < 5 ppm).

Instrumentation. Reaction experimental design was aided by the use of Accelrys Library Studio. On completion of solution dosing the plates were covered by a perfluoroalkoxy alkane (PFA) mat (Analytical Sales, Cat. No. 96967 and 24261), followed by two silicon rubber mats (Analytical Sales, Cat. No. 96965 and 24262), and an aluminum cover which was tightly and evenly sealed by 9 screws. Reactions were analyzed using a Waters Acquity UPLC. Column: Acquity UPLC BEH C18 1.7 μm 2.1 × 50 mm (Part No. 186002350), pH 3.5 Stock Solution: 12.6 g ammonium formate + 7.9 mL formic acid to 1 L water, Mobile Phase A: 40 mL pH 3.5 stock solution + 3960 mL Water, Mobile Phase B: 40 mL pH 3.5 stock solution + 360 mL Water + 3600 mL MeCN, Strong Wash: 300 mL IPA + 693 mL MeCN + 7 mL pH 3.5 stock solution, Weak Wash: 99 mL MeCN + 891 mL Water + 10 mL pH 3.5 stock solution. The instrument was equipped with an SQD detector with electrospray S4 ionization (ESI) source in the positive mode. High throughput data analysis was performed using Virscidian Analytical Studio software.

Proton nuclear magnetic resonance (^{1}H NMR) spectra were recorded on a Bruker 500 MHz VANCE spectrometer. Proton chemical shifts are reported in parts per million downfield from tetramethylsilane and are referenced to residual protium in the NMR solvent ($CHCl_3$ = δ 7.26 ppm). Carbon nuclear magnetic resonance (^{13}C NMR) spectra were recorded on a Bruker 500 AVANCE spectrometer (125 MHz). Chemical shifts for carbon are reported in parts per million downfield from tetramethylsilane and are referenced to the carbon resonances of the solvent residual peak ($CDCl_3$ = δ 77.16 ppm). Phosphorus nuclear magnetic resonance (^{31}P NMR) spectra were recorded on a Bruker 500 AVANCE spectrometer (203 MHz). Fluorine nuclear magnetic resonance (^{19}F NMR) spectra were recorded on a Bruker 500 AVANCE spectrometer (470 MHz).

High-resolution mass spectrometry was performed on an Agilent 6220 LC/MS using electrospray ionization time-of-flight (ESI-TOF), FT-IR spectra were recorded on a Perkin-Elmer Paragon 500 and are reported in terms of frequency of absorption (cm^{-1}).

Nanomolar Scale Chemistry Experiment. Nanoscale reactions (100 nmol) were run using Corning 1536-well plates (Corning Echo qualified, Cat. No. 3730, Cyclic OlefinCopolymer COC, 12.5 μL-wells, flat bottom, clear) as reaction plates, and typically with Advantage 384-well plates (Analytical Sales, Cat. No. 38120, polypropylene, 120 μL-wells, flat bottom, clear) used as solution source plates for stock solutions and for analytical plates on UPLC-MS or HPLC-MS equipment. Dosing of reaction components into the 384- and 1536-well plates was accomplished in the glovebox using a Mosquito HTS liquid handling robot (TTP Labtech, 4.5 mm pitch tip spool) with no special modifications, and using the TTP Labtech native software. Upon dosing, the 1536-well plates were covered by a perfluoroalkoxy alkane (PFA) mat (Analytical Sales, Cat. No. 96981), followed by a silicon rubber mat (Analytical Sales, Cat. No. 96982) and then secured with a clamp (Arctic White, Cat. No. AWSC-051001).

II. Ultra High Throughput Screening

Reaction Setup

The following solutions were prepared in DMSO: catalyst (0.05 M), aryl halide (0.50 M), toluidine (0.50 M), additive (0.50 M), and base (0.75 M). These solutions were added to a 384-well source plate (80 μL per well). The Mosquito HTS liquid handling robot was used to dose each of these solutions (200 nL each) into a 1536-well plate. The plate was sealed and heated to 60 °C. After 16 h, the plate was opened and the Mosquito was used to add internal standard to each well (3 μL of 0.0025 M di-tert-butylbiphenyl solution in DMSO). At that point, aliquots were sampled into 384-well plates and analyzed by UPLC.

Plate Layout

Below is a summary (Fig. S1) of all the aryl halides, Pd catalysts, bases and additives used to dose the 1536-well plates as described in the general section. The three bases were chosen due to the ease of delivery by the HTE robot since they are liquids at room temperature. Table S1 and Table S2 describe the components of each plate by rows and columns.

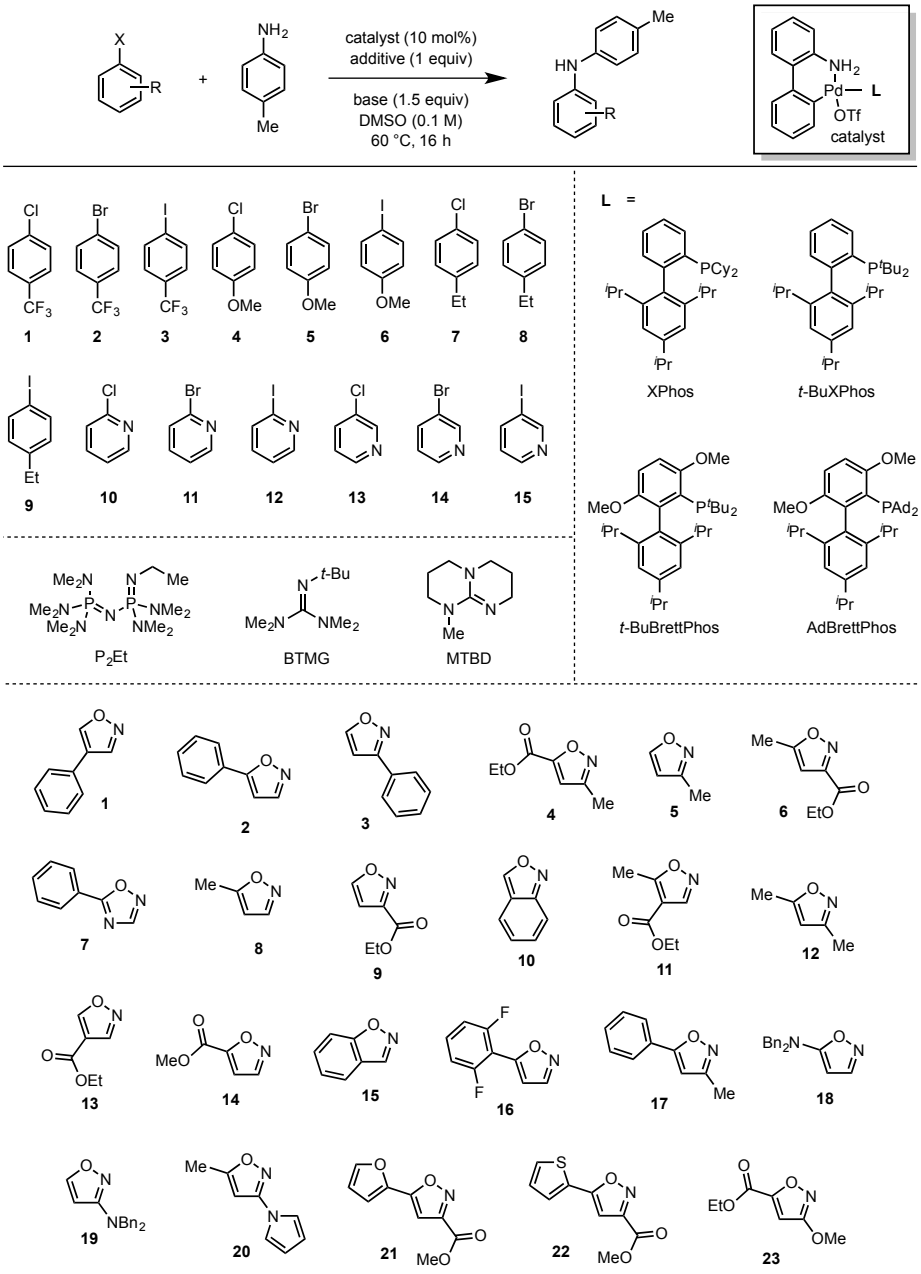


Figure S1. All reaction components.

Table S1. Layout of each plate by row.

Row	Ligand	Additive (Plate 1)	Additive (Plate 2)	Additive (Plate 3)
1	XPhos	none	**8**	**23**
2	XPhos	**2**	**10**	**17**
3	XPhos	**4**	**12**	**19**
4	XPhos	**6**	**14**	**21**
5	*t*-BuXPhos	none	**8**	**23**
6	*t*-BuXPhos	**2**	**10**	**17**
7	*t*-BuXPhos	**4**	**12**	**19**
8	*t*-BuXPhos	**6**	**14**	**21**
9	*t*-BuBrettPhos	none	**8**	**23**
10	*t*-BuBrettPhos	**2**	**10**	**17**
11	*t*-BuBrettPhos	**4**	**12**	**19**
12	*t*-BuBrettPhos	**6**	**14**	**21**
13	AdBrettPhos	none	**8**	**23**
14	AdBrettPhos	**2**	**10**	**17**
15	AdBrettPhos	**4**	**12**	**19**
16	AdBrettPhos	**6**	**14**	**21**
17	XPhos	**1**	**9**	**16**
18	XPhos	**3**	**11**	**18**
19	XPhos	**5**	**13**	**20**
20	XPhos	**7**	**15**	**22**
21	*t*-BuXPhos	**1**	**9**	**16**
22	*t*-BuXPhos	**3**	**11**	**18**
23	*t*-BuXPhos	**5**	**13**	**20**
24	*t*-BuXPhos	**7**	**15**	**22**
25	*t*-BuBrettPhos	**1**	**9**	**16**
26	*t*-BuBrettPhos	**3**	**11**	**18**
27	*t*-BuBrettPhos	**5**	**13**	**20**
28	*t*-BuBrettPhos	**7**	**15**	**22**
29	AdBrettPhos	**1**	**9**	**16**
30	AdBrettPhos	**3**	**11**	**18**
31	AdBrettPhos	**5**	**13**	**20**
32	AdBrettPhos	**7**	**15**	**22**

Table S2. Layout of each plate by column.

Column	Base	Aryl Halide	Column	Base	Aryl Halide
1	P_2Et	**1**	25	BTMG	**9**
2	P_2Et	**2**	26	BTMG	**10**
3	P_2Et	**3**	27	BTMG	**11**
4	P_2Et	**4**	28	BTMG	**12**
5	P_2Et	**5**	29	BTMG	**13**
6	P_2Et	**6**	30	BTMG	**14**
7	P_2Et	**7**	31	BTMG	**15**
8	P_2Et	**8**	32	BTMG	none
9	P_2Et	**9**	33	MTBD	**1**
10	P_2Et	**10**	34	MTBD	**2**
11	P_2Et	**11**	35	MTBD	**3**
12	P_2Et	**12**	36	MTBD	**4**
13	P_2Et	**13**	37	MTBD	**5**
14	P_2Et	**14**	38	MTBD	**6**
15	P_2Et	**15**	39	MTBD	**7**
16	P_2Et	none	40	MTBD	**8**
17	BTMG	**1**	41	MTBD	**9**
18	BTMG	**2**	42	MTBD	**10**
19	BTMG	**3**	43	MTBD	**11**
20	BTMG	**4**	44	MTBD	**12**
21	BTMG	**5**	45	MTBD	**13**
22	BTMG	**6**	46	MTBD	**14**
23	BTMG	**7**	47	MTBD	**15**
24	BTMG	**8**	48	MTBD	none

Sample Yield Determination

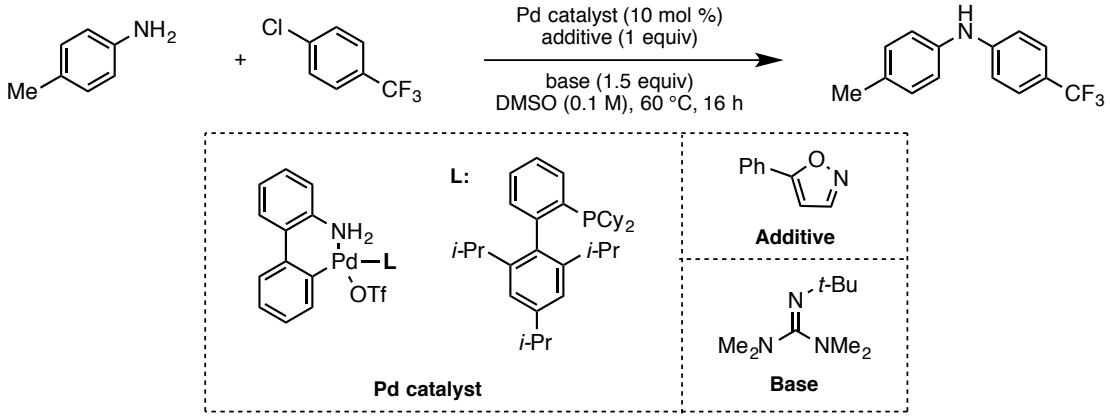
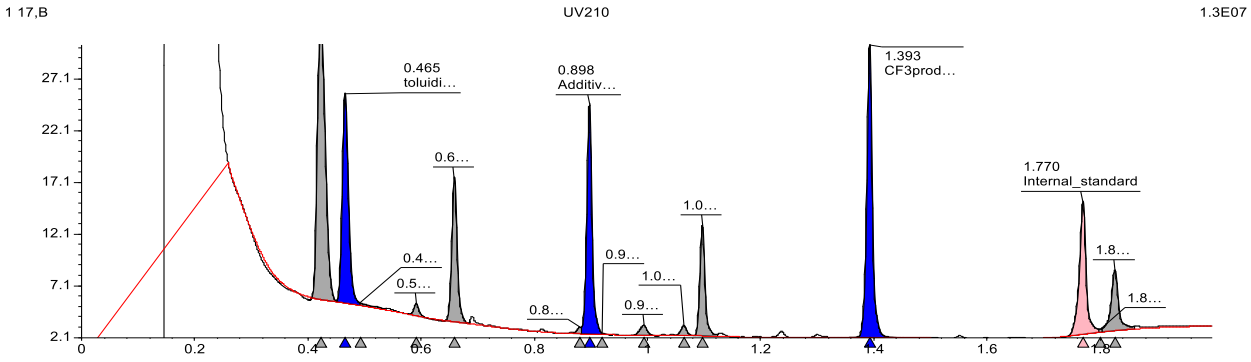


Figure S2. Reaction summary for Plate 1, Row 2, Column 17.

The reaction above (Fig. S2) was set up in 1536-well Plate 1 following the HTE protocol described in the general section. After 16 h, the reaction was analyzed by UPLC and the yield determined using di-*tert*-butylbiphenyl as an internal standard (Fig. S3 and Table S3). Whenever possible, remaining starting material and additive were also recorded. However, occasionally the additive peak overlapped with nearby peaks making it difficult to quantify. As a result, the amount of additive remaining was not used to model trends in reaction yield.



Location	toluidine Rf (min)	toluidine Area Abs
17:B	0.465	34740.88184
UV210	CF3product Rf (min)	CF3product Area Abs
	1.393	43359.57595
	Additive2 Rf (min)	Additive2 Area Abs
	0.898	31989.22293
	Internal_standard Rf (min)	Internal_standard Area Abs
	1.770	24180.02388

Figure S3. Chromatographic data for Plate 1, Row 2, Column 17.

Table S3. Retention times for all products.

Product	**Rf (min)**
CF3	1.39
OMe	1.22
Et	1.42
2-pyr	0.75
3-pyr	0.72

Plate Heatmaps

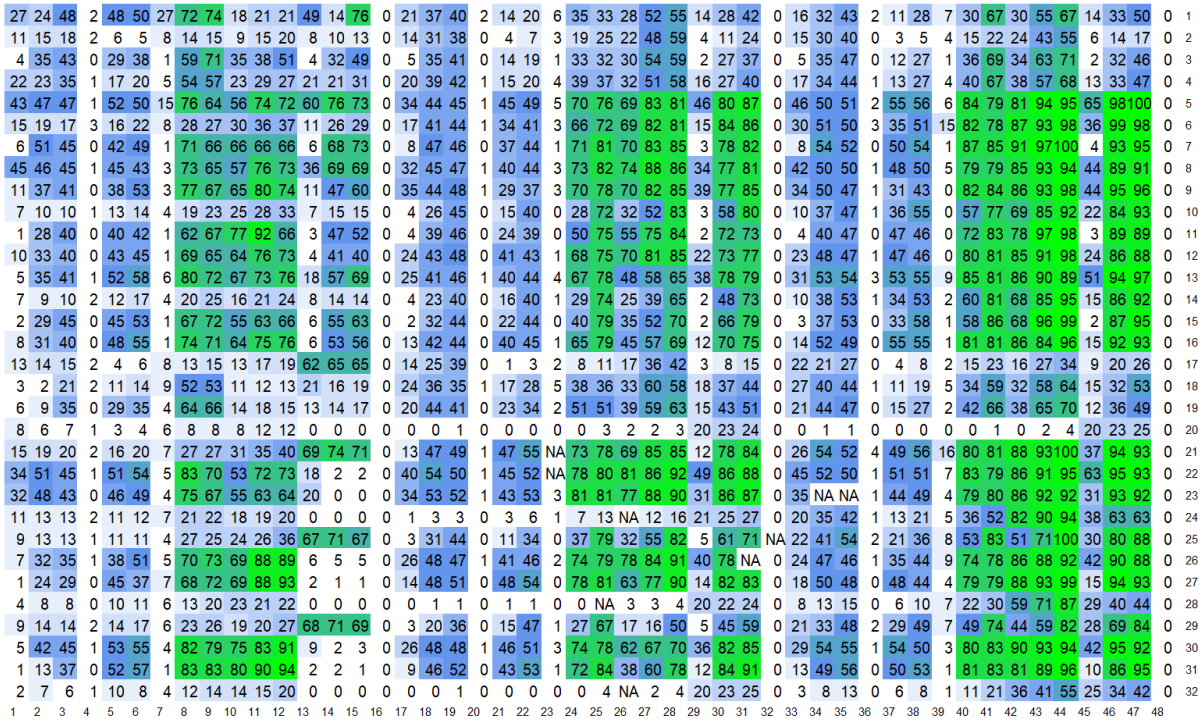


Figure S4. Plate 1 yields.

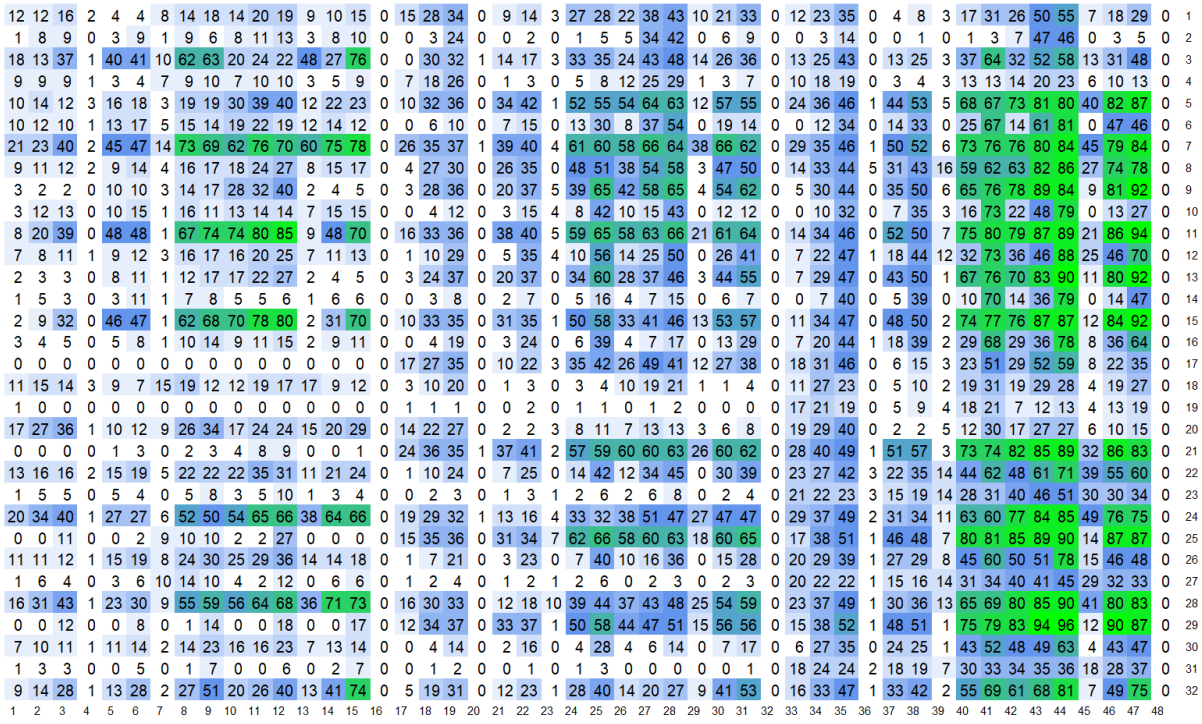


Figure S5. Plate 2 yields.

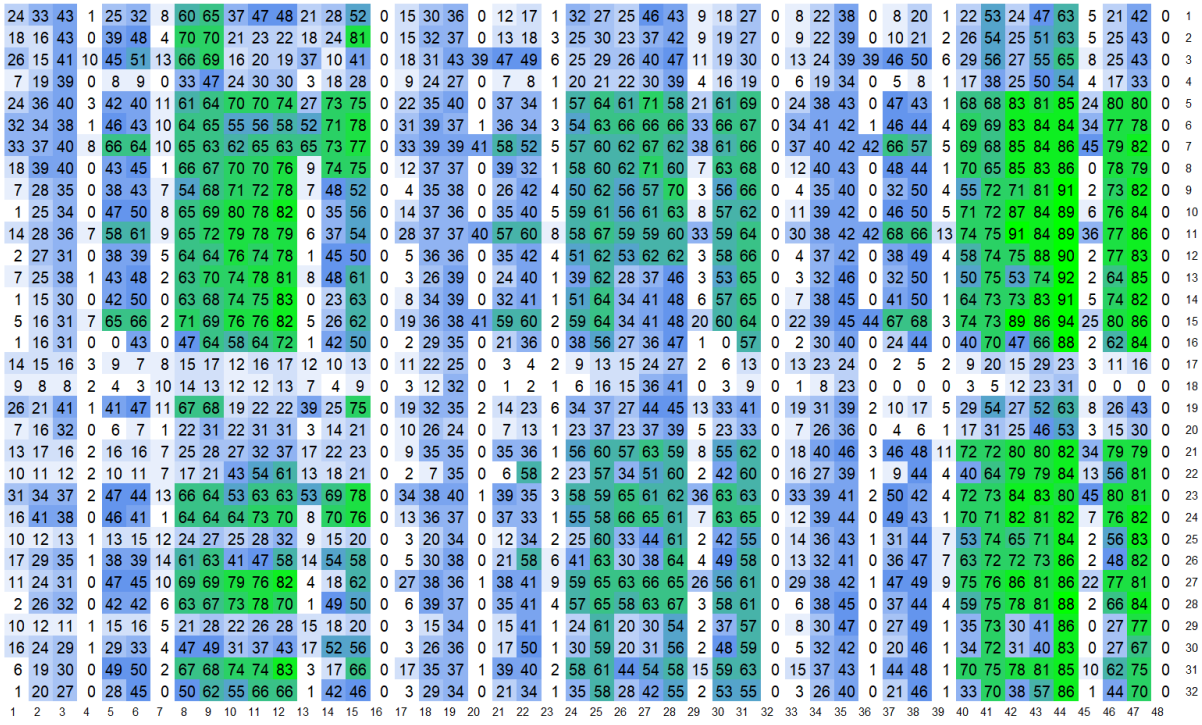


Figure S6. Plate 3 yields.

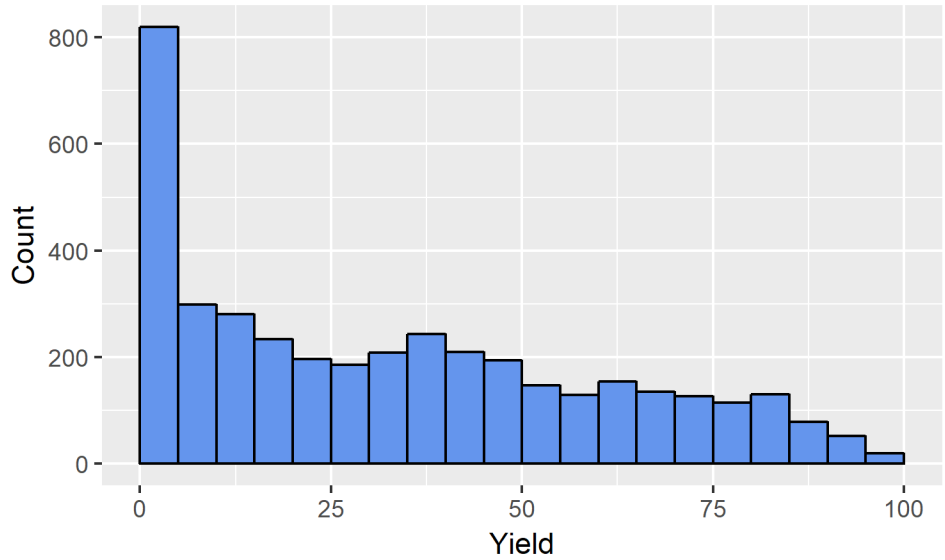


Figure S7. Reaction yields used for modeling (control reactions and reactions containing additive **7** removed).

III. Programming Details

Code, data, and instructions can be found at https://github.com/doylelab/rxnpredict.

Instructions for Using `rxnpredict`

1. Download and install the following programs:
    - Spartan '14 V1.1.4
    - Python 3 (The anaconda distribution is recommended, as it has packages required for the software to run: Download at https://www.anaconda.com/download/)
    - R (Download at https://cran.r-project.org/mirrors.html – choose any mirror link)
    - R Studio (Download at https://www.rstudio.com/products/rstudio/download/)
    - Sublime Text 3 (Download at https://www.sublimetext.com/3)
2. Add Anaconda as a PATH variable so that Python will execute the scripts within Sublime Text 3.
    - Navigate to "This PC" in File Explorer
    - Right click "This PC" → Properties → Advanced system settings → Environment Variables...
    - In User variables, click "Path" variable → Edit → New → type "path\to\Anaconda3" (no quotes – e.g., C:\Users\Derek\Anaconda3)
3. Go to https://github.com/doylelab/rxnpredict. On right, click "Clone or Download" and then "Download Zip". This will download a local copy of the repository (folder) to your computer.
4. All of the molecules whose properties will be used for modeling must first be drawn in Spartan:
    - Use the Spartan GUI to draw the molecules, saving them in the spartan_molecules folder (within the rxnpredict folder).
    - Be sure to label any shared atoms within a substrate class (ligand, base, etc.) with a "*". You can do so by right clicking an atom, then click "Properties". Change the label text at the bottom of the dialog box (e.g., *C1).
    - Save the molecules in both .spardir format (for future editing) and .spinput format (this is what the program uses).
5. Modify the python scripts:
    - In setup.py, change the value of spartan_path (line 16 only) to the path of the Spartan14v114.exe file. Be sure to use \\ between folder names.
    - In main.py, describe the 2D layout of the plates you have run (line 10 onwards). Helpful syntax:
        - `plate_name = Plate(x,y)` where x is the number of rows and y is the number of columns.
        - `plate_name.fillRow([list], 'substrate_class', 'molecule_name')` where 'molecule_name' corresponds to the name of that molecule's .spinput file. Replace `fillRow` with `fillColumn` to populate columns instead. Note: If your plate design does not conform to one

molecule per row/column, you can modify the Plate's dimensions accordingly. If necessary, an Nx1 plate can specify the components of each reaction individually.

- After the plates are filled (all "cells" of the plates must be populated with the same kinds of substrate_class), insert the following lines (where the plate names match the ones you have created):

```
setup.export_reactions([plate1,plate2,plate3])
setup.export_for_pca([plate1,plate2,plate3])
```

- Run main.py (ctrl + B in Sublime Text) to create R\output_table.csv. This table consists of one row per reaction and one column per descriptor per reaction component. For example, a 2000-reaction screen with 20 base descriptors, 50 ligand descriptors, and 30 additive descriptors would generate a table with 2000 rows and 100 columns in R\output_table.csv.

6. Run the analysis_template.R file:
   - First, save yield data in Excel in a single column in a file named yields.csv in the rxnpredict folder (note that the file will need to be saved in .csv format). Reactions without yield data due to analytical or other issues should be coded as NA.
   - Open analysis_template.R in R Studio and modify the working directory to the location of the rxnpredict folder (line 9). [Note: Before running the R code on a Mac, change the file locations such that folders are separated by "/" instead of "\\".] Run the code by pressing ctrl + A and then ctrl + enter, which will perform the following steps:
     - Loading and scaling the output data generated by the python script.
     - Merging the yield data from yields.csv onto the dataset.
     - Splitting the data into training (70%) and test (30%) sets.
     - Training a random forest model using the training set.
     - Plotting a calibration plot of the model using the test set.
     - Calculating R^2 and RMSE values for the model using the test set.
     - Generating a variable importance plot for the model.
   - In the R Studio console, the test set R^2 and RMSE values should be printed in black text. A calibration plot and variable importance plot should be located in the R\plots folder.

How the Program Works

This section provides more detailed information about how the Python scripts within the `rxnpredict` program operate. After completing steps 1-3 above, the user draws the chemical structure of each reagent (e.g., aryl halides, ligands, bases) in the Spartan GUI. These structures should be saved as .spinput files in the spartan_molecules folder, which is inside the rxnpredict folder. Be sure to label any shared atoms within a substrate class (ligand, base, etc.) with a "*". You can do so by right clicking an atom, clicking "Properties", and changing the label text at the bottom of the dialog box. The atoms which were labeled in our study are shown below.

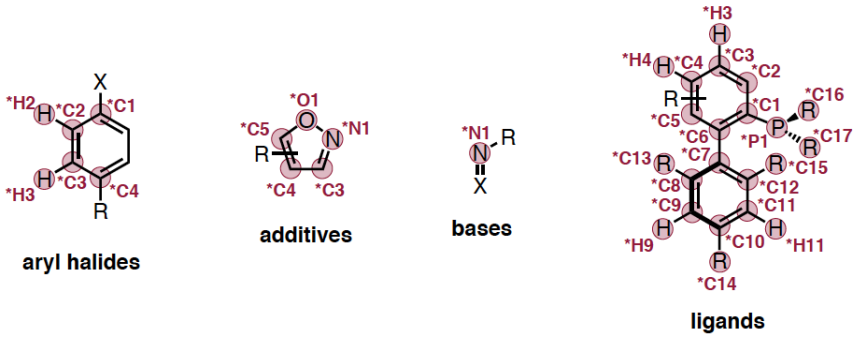


Figure S8. Shared atoms for each substrate class.

The user must modify the main.py and setup.py files as described above in the instructions. Upon running setup.py (ctrl + B in Sublime Text), the program will compute and extract molecular, atomic, and vibrational descriptors and construct an output table which can be used to build models in R.

```
molecule.spardir
├── _spartandir
└── M0001
    ├── _spartan
    └── input
```

Figure S9. Necessary folder structure for Spartan calculations.

The sections below describe in greater detail how this process works. When setup.py is run, the program first extracts the Cartesian coordinates, atomic labels, and Hessian matrix from the molecule.spinput file. After the directory structure is created (Fig. S9), the M0001/input file is populated with this molecular information. To the input file is also added computational details, such as convergence parameters and descriptor output options. The input file for 3-methylisoxazole is shown as a representative example (Fig. S10).

3-methylisoxazole/M0001/input

```
C OPT B3LYP 6-31G* FINDBEST=MMFF FREQ NMR POLAR
C CONVERGE SCFTOLERANCE=VERYHIGH GRADIENTTOLERANCE=0.000005
DISTANCETOLERANCE=0.00002 BIGGRID
M0001
0 1
 8   0.134073674   0.416507062   1.125571397
 7   1.104325733   0.214400283   2.092079714
 6   0.443436054   0.061231247   3.244640007
 6  -0.951312271   0.159864267   3.050918356
 6  -1.102251887   0.384220987   1.699635900
 1  -1.738133734   0.079700106   3.786216269
 6   1.185939110  -0.176783187   4.502194474
 1   1.778020516  -1.094928256   4.431891440
 1   0.501966682  -0.279909369   5.350707876
 1   1.863000613   0.657182302   4.713339677
 1  -1.973785727   0.530290562   1.051323904
ENDCART
ATOMLABELS
"*O1"
"*N1"
"*C3"
"*C4"
"*C5"
"H8"
"C6"
"H3"
"H5"
"H6"
"H4"
ENDATOMLABELS
HESSIAN
    8    6    2    2    2   13    1   13   13   13   13
    1    2    1
    1    5    1
    2    3    2
    3    4    1
    4    6    1
    4    5    2
    7    8    1
    7    9    1
    7   10    1
    3    7    1
    5   11    1
ENDHESS
NOCONFORMER
BEGINPROPIN
C DOQSAR PRINTMO THERMO PRINTFREQ PRINTCHG
C QSAR NBO DIPOLE POP PROPPRINTLEV=2 PROPRERUN
 PRINTVIBCOORDS PROP:IGNORE_WARN
ENDPROPIN
BEGINPREFERENCES
 MM:CONF_SELECTION_RULE=2
ENDPREFERENCES
```

Figure S10. Input file for 3-methylisoxazole.

Once the input file is created, the program submits the Spartan DFT calculations *via* the command line (B3LYP/6-31G*). During these calculations, Spartan performs a geometry

optimization prior to the DFT, frequency, and property calculations. As currently written, the command line submissions do not work on a Mac due to the different syntax between these systems. During the Spartan calculation, a number of files are generated within this folder structure, including a file named "output." This file is placed in the molecule's M0001 folder and contains all relevant computational outputs. At this point, molecular, atomic, and vibrational descriptors are extracted.

```
                    Cartesian Coordinates (Angstroms)
      Atom            X             Y             Z
    ---------    ------------ ------------ ------------

  1 O  *O1          1.3849524    0.7620180    0.0000000
  2 N  *N1          0.0291077    1.1133151    0.0000000
  3 C  *C3         -0.6341170   -0.0237965    0.0000000
  4 C  *C4          0.2517957   -1.1447768    0.0000000
  5 C  *C5          1.4863681   -0.5776663    0.0000000
  6 H  H8          -0.0038756   -2.1939467    0.0000000
  7 C  C6          -2.1364791   -0.0275569    0.0000000
  8 H  H3          -2.5285919    0.4867135    0.8834114
  9 H  H5          -2.5195350   -1.0514516    0.0000000
 10 H  H6          -2.5285919    0.4867135   -0.8834114
 11 H  H4           2.4918152   -0.9746002    0.0000000

  Point Group = CS  Order =  1 Nsymop =  2

Molecular descriptors:

Molecular volume:        88.98 (Ang**3)
Surface area:           109.97 (Ang**2)

Ovality:                  1.141

Atomic weight:           83.090 g

E(HOMO):                 -0.2612
E(LUMO):                 -0.0106

Electronegativity:        0.14
Hardness:                 0.13
```

Figure S11. Excerpt from 3-methylisoxazole.spardir\M0001\output file showing atomic positions for optimized geometry and some molecular descriptors.

Molecular descriptors were extracted using appropriate keywords in the output file, such as "Molecular volume:" (see Fig. S11). The extracted molecular descriptors include molecular volume, surface area, ovality, molecular weight, E_{HOMO}, E_{LUMO}, electronegativity, hardness, and dipole moment.

Atomic descriptors were also parsed from the output file. First, the shared atoms were determined for each reagent class (e.g., the shared atoms for the additives were *O1, *N1, *C3, *C4, and *C5). These are determined by comparing atom labels between each molecule within a reagent class (all of the aryl halides, for instance). All of the atomic labels which contain "*" and are present in every molecule are considered to be shared atoms. This procedure underscores the importance of the user labeling the atoms appropriately. For each of the shared atoms, the calculated electrostatic charge and calculated NMR shift were extracted using keywords in the output file.

To extract information related to shared molecular vibrations, we first had to determine which molecular vibrations were indeed shared by all molecules in a substrate class (base, aryl halide, etc.). We used the following strategy to compare the similarity of molecular vibrations. In the output file, each molecular vibration is represented as an N×3 matrix where each row is an atom and each column is an *x*-, *y*-, or *z*-coordinate associated with that atom's movement. We first extracted this atomic movement data and subsetted it to only the shared atoms for each vibrational mode. An example of the atomic movement information is shown for six molecular vibrations (Fig. S12).

In order to compare the atomic movement for two molecular vibrations, the same coordinate system must be used. Therefore, the atomic coordinates of each molecule's optimized geometry (not the coordinates for the molecular vibrations) was used to find the rotations in each axis which would minimize the distance between matching atoms. After testing that labeled atoms in each of the molecules were not collinear (to ensure that an unambiguous coordinate system could be defined), one of the molecules was rotated to minimize the distance between common atoms. Note that the molecules will not overlay exactly due to the geometry optimizations which have been performed on each structure. This sequence of rotations is then applied to the molecular vibrations such that they can be directly compared. After flattening the atomic movement matrix from an N×3 matrix to a vector of length 3N, the Pearson correlation coefficient (R^2) was computed between these two molecular vibrations (Fig. S13). Atomic movement weighted by atomic mass was empirically found to be a better predictor of shared vibrational modes than atomic movement alone. If not weighted, the movement of hydrogen atoms tends to exhibit outsized importance relative to the movements of heavier atoms. Therefore, the final algorithm computes R^2 values for flattened atomic movement vectors where each vector is multiplied by that atom's atomic mass.

Normal Modes and Vibrational Frequencies (cm-1)

	-129.73			273.92			340.32		
	A"			A"			A'		
	X	Y	Z	X	Y	Z	X	Y	Z
*O1	0.000	0.000	-0.016	0.000	0.000	-0.071	0.051	-0.027	0.000
*N1	0.000	0.000	-0.044	0.000	0.000	0.092	0.059	0.080	0.000
*C3	0.000	0.000	-0.010	0.000	0.000	0.138	0.000	0.108	0.000
*C4	0.000	0.000	0.034	0.000	0.000	0.093	-0.078	0.061	0.000
*C5	0.000	0.000	0.027	0.000	0.000	-0.085	-0.042	-0.036	0.000
H8	0.000	0.000	0.060	0.000	0.000	0.113	-0.179	0.086	0.000
C6	0.000	0.000	0.009	0.000	0.000	-0.102	0.004	-0.131	0.000
H3	0.009	-0.484	0.297	-0.192	-0.037	-0.166	-0.121	-0.231	0.003
H5	0.000	0.000	-0.534	0.000	0.000	-0.234	0.246	-0.223	0.000
H6	-0.009	0.484	0.297	0.192	0.037	-0.166	-0.121	-0.231	-0.003
H4	0.000	0.000	0.041	0.000	0.000	-0.227	-0.071	-0.109	0.000

	609.83			659.19			664.35		
	A"			A'			A'		
	X	Y	Z	X	Y	Z	X	Y	Z
*O1	0.000	0.000	-0.149	0.090	-0.004	0.000	0.000	0.000	-0.020
*N1	0.000	0.000	0.115	0.036	-0.023	0.000	0.000	0.000	0.120
*C3	0.000	0.000	-0.003	-0.059	0.003	0.000	0.000	0.000	-0.188
*C4	0.000	0.000	-0.088	0.057	0.036	0.000	0.000	0.000	0.113
*C5	0.000	0.000	0.133	0.082	0.001	0.000	0.000	0.000	-0.042
H8	0.000	0.000	-0.041	0.138	0.014	0.000	0.000	0.000	0.244
C6	0.000	0.000	-0.004	-0.206	-0.001	0.000	0.000	0.000	-0.025
H3	-0.007	-0.007	-0.003	-0.221	-0.015	0.002	0.212	-0.004	0.072
H5	0.000	0.000	-0.014	-0.187	-0.010	0.000	0.000	0.000	0.081
H6	0.007	0.007	-0.003	-0.221	-0.015	-0.002	-0.213	0.004	0.072
H4	0.000	0.000	0.383	0.066	-0.047	0.000	0.000	0.000	-0.127

Figure S12. Excerpt from 3-methylisoxazole.spardir\M0001\output file showing atomic movement for six molecular vibrations.

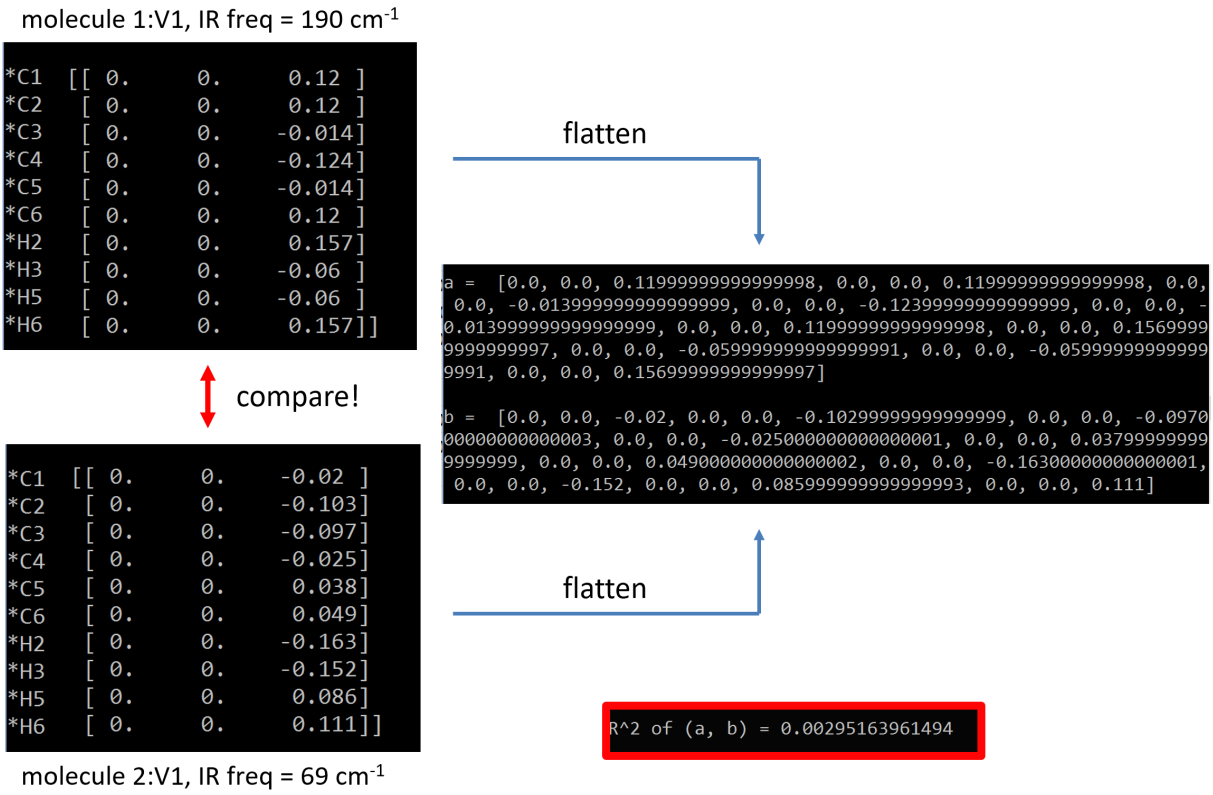


Figure S13. Comparison of two molecular vibrations using a Pearson correlation.

Computing the Pearson coefficient for every molecular vibration for a pair of molecules creates a correlation matrix (Fig. S14 and Fig. S15). For instance, if molecule A has 30 vibrational modes and molecule B has 45 vibrational modes, their correlation matrix would have 30 rows and 45 columns. Using a correlation matrix, shared molecular vibrations were determined using the following criteria:

- R^2 value must be the highest in both its row and its column.
- R^2 value must be greater than 0.50.
- Frequency of the vibration must be greater than 500 cm^{-1} (calculated frequencies below this threshold are not considered reliable).

Using one of our generated correlation matrices, we found that pairs of molecular vibrations determined using these criteria identified the same vibrational modes in all cases. Using lower thresholds for R^2 value resulted in vibrations being picked out as "shared" which were not the same. Using higher R^2 thresholds did not improve matching accuracy.

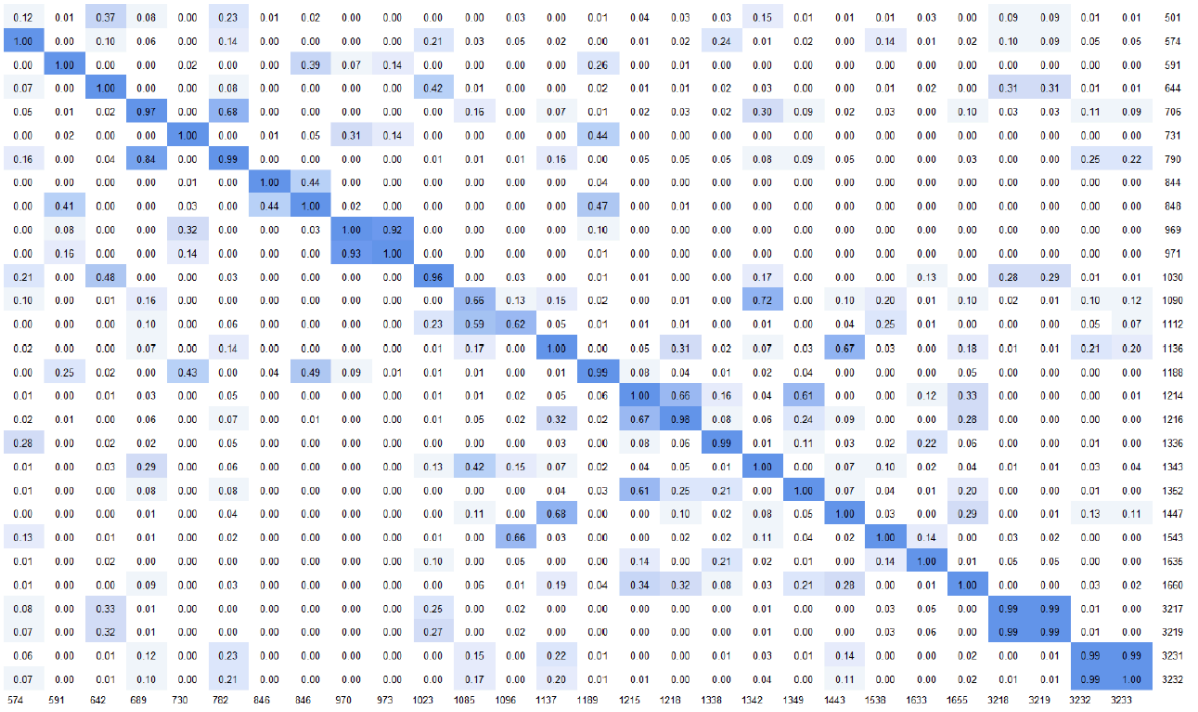


Figure S14: Correlation matrix for the vibrational modes of aryl halides **1** and **2**.

Figure S15: Correlation matrix for the vibrational modes of additives **1** and **2**.

Figure S16 shows an example of which molecular vibrations would be considered shared based on a single correlation matrix. Based on the criteria above, a list of shared vibrational modes is generated. Correlation matrices were generated between the first molecule and every remaining molecule within a reagent class. For instance, since 15 aryl halides were used in screening, 14 correlation matrices were generated for aryl halide vibrational modes (between aryl halides **1** and **2**, **1** and **3**, **1** and **4**, and so on). If every molecule had a vibration that matched a vibrational mode for the first molecule, that vibrational mode was considered to be conserved. In this case, the vibration's frequency and intensity was extracted for each molecule and these numbers would serve as the vibrational descriptors.

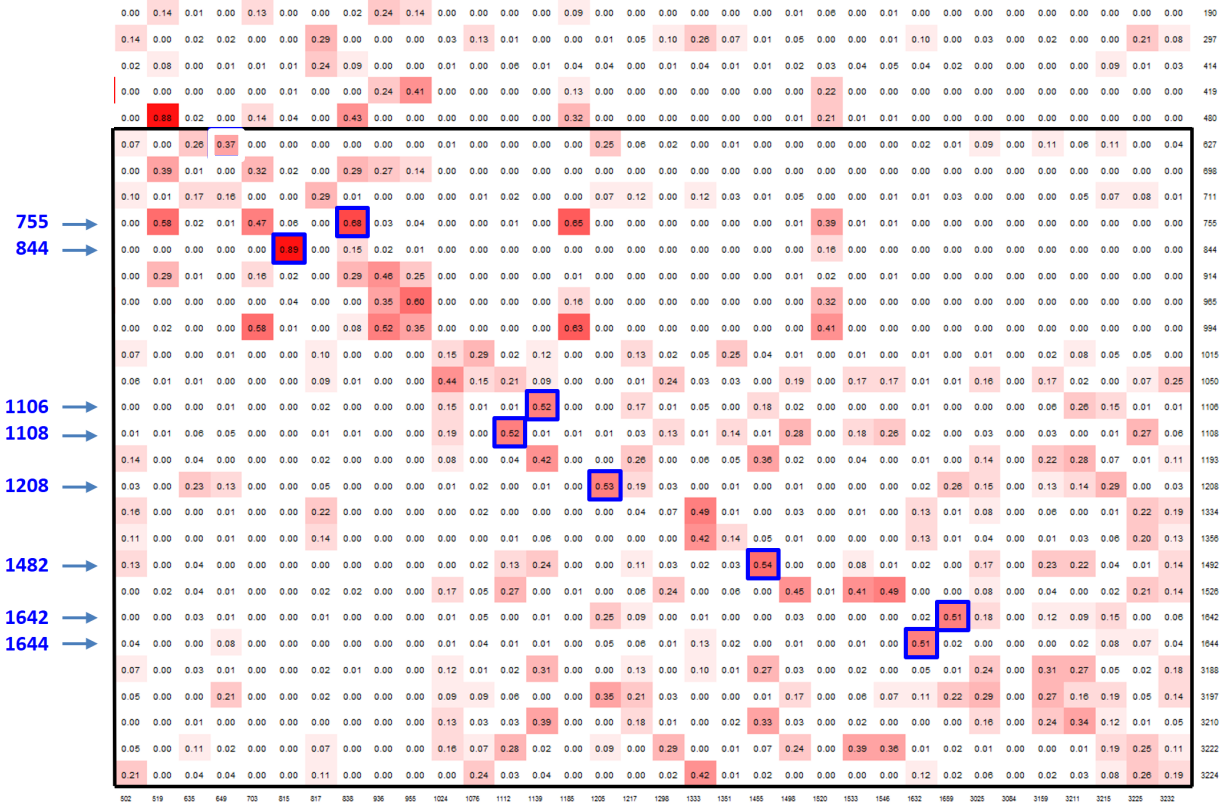


Figure S16. Shared molecular vibrations from a single correlation matrix.

Once the program extracts all of these descriptors, it creates the table shown in Table S4 (without the yield column, which is added in R). Once this yield information is added, the table can be used for modeling. In our case, a total of 120 descriptors per reaction were generated. To recap, the following descriptors are calculated, parsed, and added to the output table:

- Molecular descriptors: molecular volume, surface area, ovality, molecular weight, E_{HOMO}, E_{LUMO}, electronegativity, hardness, and dipole moment.
- Atomic descriptors (for shared atoms – ones labeled with an "*" in the .spinput files): Electrostatic charge and NMR shift
- Vibrational descriptors for shared molecular vibrations: Frequency and intensity

Table S4. Structure of output table (descriptors from Python scripts, yield is appended in R).

	Additive			Aryl Halide			Base			Ligand			
	D_1	...	D_n	D_1	...	D_n	D_1	...	D_n	D_1	...	D_n	yield
Reaction 1													
Reaction 2													
...													
Reaction N													

Additive Descriptors (*n* = 19)
E_{HOMO}, E_{LUMO}, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular Weight, Ovality, Surface Area, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, *C4 Electrostatic Charge, *C5 NMR Shift, *C5 Electrostatic Charge, *N1 Electrostatic Charge, *O1 Electrostatic Charge, V1 Frequency, V1 Intensity

Aryl Halide Descriptors (*n* = 27)
E_{HOMO}, E_{LUMO}, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular Weight, Ovality, Surface Area, *C1 NMR Shift, *C1 Electrostatic Charge, *C2 NMR Shift, *C2 Electrostatic Charge, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, *C4 Electrostatic Charge, *H2 NMR Shift, *H2 Electrostatic Charge, *H3 NMR Shift, *H3 Electrostatic Charge, V1 Frequency, V1 Intensity, V2 Frequency, V2 Intensity, V3 Frequency, V3 Intensity

Base Descriptors (*n* = 10)
E_{HOMO}, E_{LUMO}, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular Weight, Ovality, Surface Area, *N1 Electrostatic Charge

Ligand Descriptors (*n* = 64)
Dipole Moment, *C1 NMR Shift, *C1 Electrostatic Charge, *C2 NMR Shift, *C2 Electrostatic Charge, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, *C4 Electrostatic Charge, *C5 NMR Shift, *C5 Electrostatic Charge, *C6 NMR Shift, *C6 Electrostatic Charge, *C7 NMR Shift, *C7 Electrostatic Charge, *C8 NMR Shift, *C8 Electrostatic Charge, *C9 NMR Shift, *C9 Electrostatic Charge, *C10 NMR Shift, *C10 Electrostatic Charge, *C11 NMR Shift, *C11 Electrostatic Charge, *C12 NMR Shift, *C12 Electrostatic Charge, *C13 NMR Shift, *C13 Electrostatic Charge, *C14 NMR Shift, *C14 Electrostatic Charge, *C15 NMR Shift, *C15 Electrostatic Charge, *C16 NMR Shift, *C16 Electrostatic Charge, *C17 NMR Shift, *C17 Electrostatic Charge, *H11 NMR Shift, *H11 Electrostatic Charge, *H3 NMR Shift, *H3 Electrostatic Charge, *H4 NMR Shift, *H4 Electrostatic Charge, *H9 NMR Shift, *H9

Electrostatic Charge, *P1 Electrostatic Charge, V1 Frequency, V1 Intensity, V2 Frequency, V2 Intensity, V3 Frequency, V3 Intensity, V4 Frequency, V4 Intensity, V5 Frequency, V5 Intensity, V6 Frequency, V6 Intensity, V7 Frequency, V7 Intensity, V8 Frequency, V8 Intensity, V9 Frequency, V9 Intensity, V10 Frequency, V10 Intensity

Modeling

Once the output table was generated by the Python scripts, it was used to model reaction outcomes. The code for steps described below can be found in the analysis.R file at https://github.com/doylelab/rxnpredict. All of the machine learning modeling algorithms, except for the regularized linear models, were implemented using the R `caret` package. The descriptor data was centered and scaled prior to modeling using the `scale(x)` function in R. This function normalizes the descriptors by substracting the mean and dividing by the standard deviation. Seeds (`set.seed(x)`) were added to the R code to ensure reproducibility of the models and figures shown on Github.

Our first objective was to validate that the high-dimensional reaction data could be effectively modeled. At the outset of our studies, it was unclear whether it was possible to accurately model reaction outcomes in many dimensions, as the corresponding energy surfaces can be relatively shallow. The metrics used to evaluate all of the models discussed herein were root mean square error (RMSE) and the coefficient of determination (R^2) for the corresponding calibration plots. A model with good fit is characterized by a low RMSE and high R^2 value. At the outset, we examined a variety of regularized linear regression methods (*37*). These techniques seek to minimize the effects of overfitting by imposing a cost on model complexity. The least absolute shrinkage and selection operator, or LASSO, is a regression analysis method which performs both regularization and variable selection (*38*). The lasso estimator is

$$\hat{\beta}^{\text{LASSO}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\}$$

where λ_1 is a tuning parameter which controls the magnitude of regularization. For larger values of λ_1, the regression coefficients experience greater shrinkage. Ridge regression, which introduces a quadratic penalty term to the loss function, was also explored. The ridge estimate is defined by the Lagrangian form

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

In this case, λ_2 is a complexity parameter which controls the extent of shrinkage. Elastic net is a third regularization method that was evaluated. The elastic net method contains a combination of the LASSO and ridge penalties (*39*):

$$\hat{\beta}^{\text{enet}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \right\}$$

where $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. When $\alpha = 0$, elastic net becomes LASSO regression, and when $\alpha = 1$, the expression simplifies to ridge regression. As defined, α takes values between 0 and 1, representing the relative contributions of these shrinkage methods. As seen in Fig. S17, elastic net outperforms both LASSO and ridge regression. However, this improvement is relatively small, achieving a RMSE of 15.58 and an R^2 value of 0.8179.

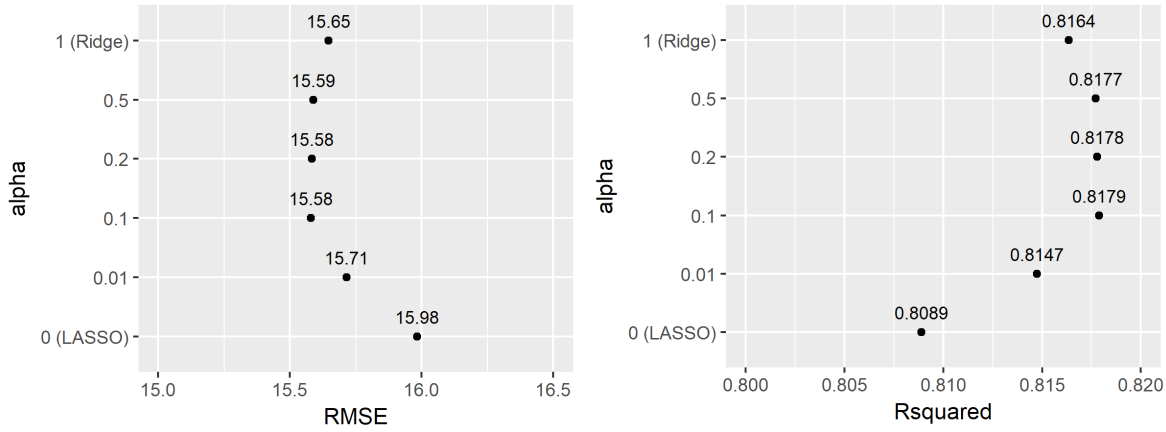


Figure S17. Test set performance using regularized regression models.

Variable reduction was also explored as a means for generating an effective linear model. To reduce the number of descriptors, we sought to remove correlated descriptors, as these undermine the generation of effective linear models. We first generated correlation plots for each class of substrates (Fig. S18). These plots show the extent to which different descriptors are correlated with each other. Each square shows the covariance (r) between two lists of descriptors (imagine plotting the *C1 NMR shift vs. *C3 NMR shift for the 15 aryl halides). The descriptors shown below represent those remaining when covariances between –0.5 and 0.5 are kept (the algorithm for descriptor removal is stepwise). This corresponds to r^2 values of less than 0.25. Correlation plots with these remaining descriptors is shown in Fig. S19.

Additive
*C3 Electrostatic Charge, *C5 NMR Shift, *O1 Electrostatic Charge, V1 Frequency, V1 Intensity, Dipole Moment, Electronegativity, Hardness, Ovality

Aryl Halide
*C1 NMR Shift, *C3 NMR Shift, *C3 Electrostatic Charge, *C4 NMR Shift, Dipole Moment, Electronegativity, Ovality

Base
E_{LUMO}, Ovality

Ligand
*P1 Electrostatic Charge, V1 Intensity, V6 Frequency

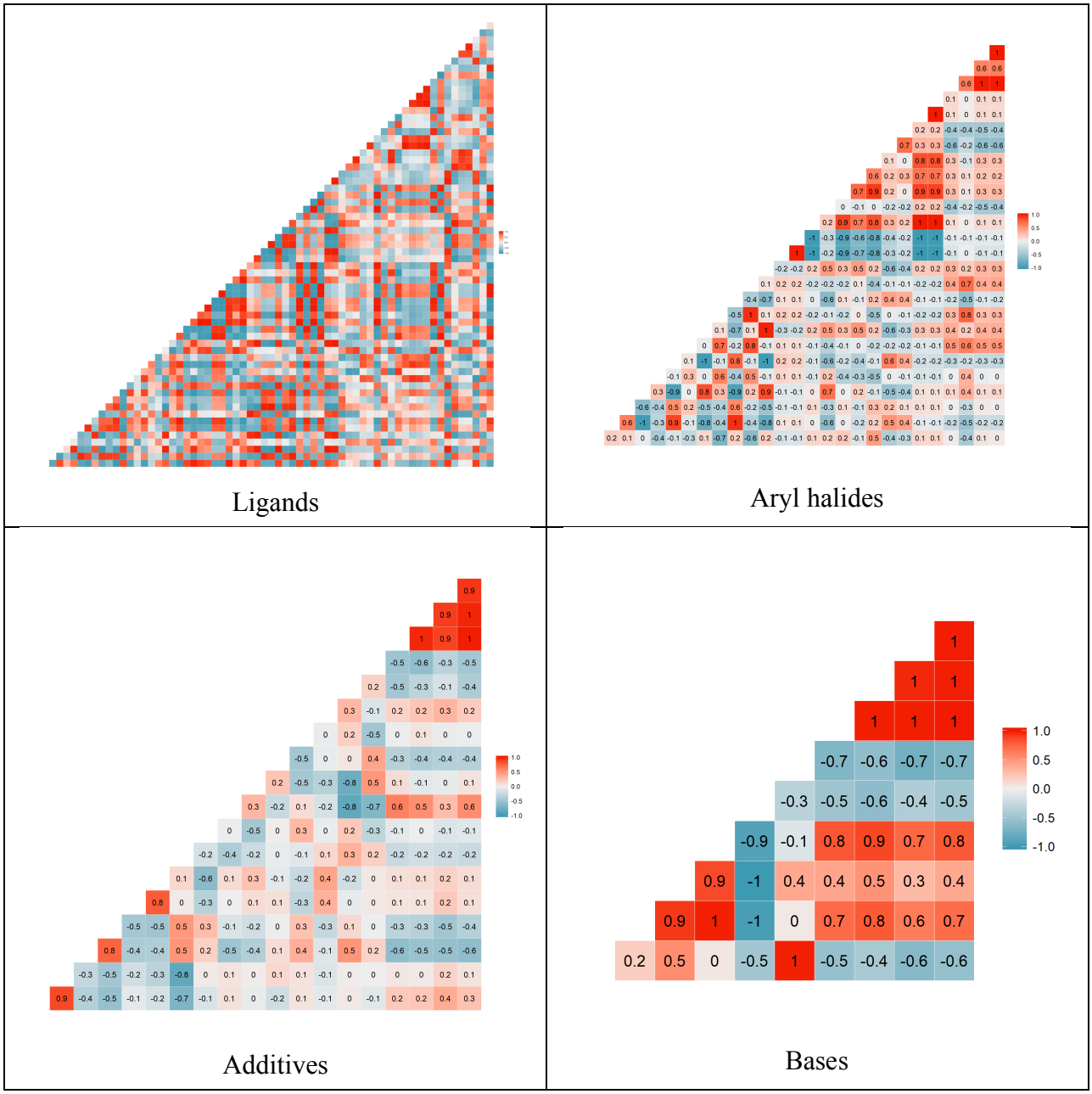


Figure S18. Correlation plots for descriptors within each substrate class.

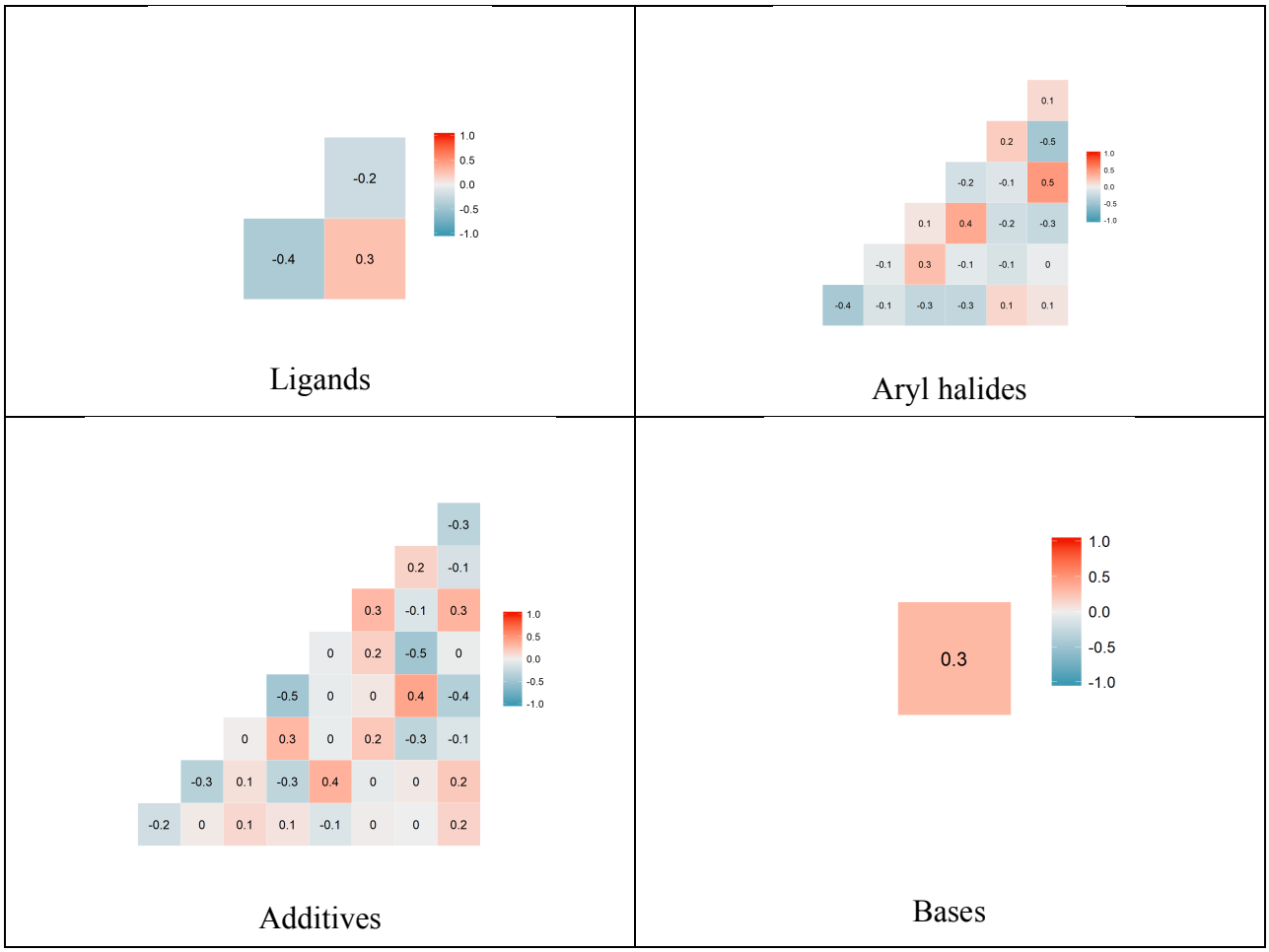


Figure S19. Correlation plots for descriptors with correlated descriptors removed.

When a linear model was constructed using the shown reduced descriptor set ($r^2 < 0.25$), the calibration plot shown in Fig. S20 was obtained. The generation of this model employed a 70/30 split of training and test data. This model had an R^2 value of 0.696 and a RMSE of 19.44. Notably, a variety of different r^2 thresholds were tried, but none of the reduced descriptors sets produced a better linear model than the parent linear model using all descriptors.

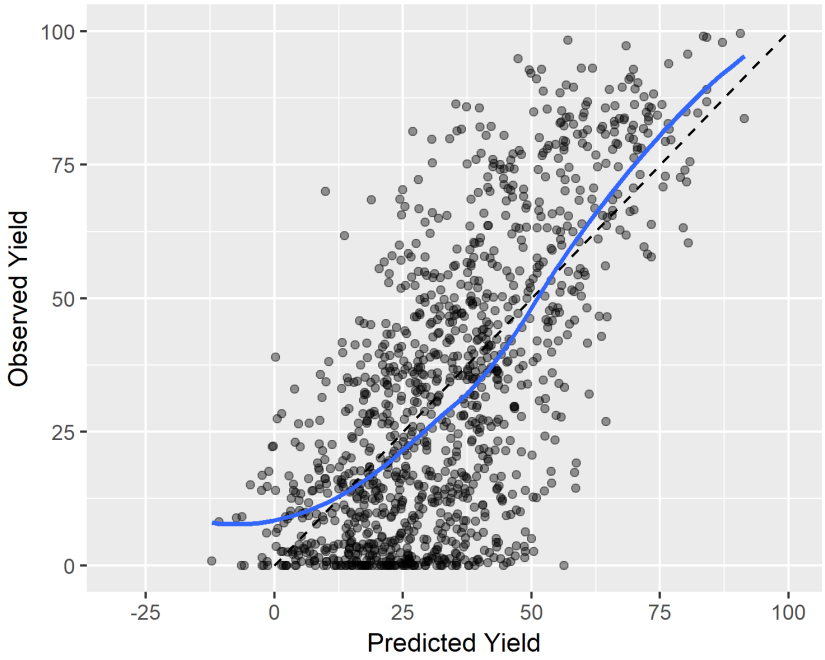


Figure S20. Linear model using reduced set of descriptors.

At this point, a number of supervised machine learning algorithms were used to model the experimental data. These models were used to identify the relationship between reaction conditions (i.e., molecular descriptors for aryl halides, precatalysts, bases, and additives) and the reaction outcome (i.e., product yield). We performed a 70/30 split of training and test data, and performed 10-fold cross-validation on the training data to measure each model's generalizability to an independent dataset. These analyses were performed in R using the `caret` package. Calibration plots were generated to evaluate the success of different model types (Fig. S21). For each point, the *x*-value is the predicted yield based on the trained regression model and the *y*-value is the observed yield which was experimentally obtained. Loess best-fit curves (solid) and the $y = x$ line (dashed) are included in each plot. Ideally, the best-fit line for these points should have a slope approximating one and an intercept near zero. Finally, true error estimation was performed using the holdout test data.

In order to perform both model selection and true error estimation, it is standard to perform nested cross-validation. Below are the training performances of each ML algorithm when cross-validation was performed on the training set (70% of all reactions). Based on the cross-validation RMSE and R^2 values shown below, random forest clearly emerges as being the most promising algorithm. The calibration plots discussed below and in the manuscript were generated using the test set data so as to show the generalizability of each method within the rest

of the dataset. Also, calibration plots are more intuitive for the test set than in the reactions used for cross-validation, which are employed for both training and testing.

```
> getTrainPerf(lmFit)
  TrainRMSE TrainRsquared method
1  14.78928     0.7105538     lm

> getTrainPerf(knnFit)
  TrainRMSE TrainRsquared method
1  16.81863     0.6300943    knn

> getTrainPerf(svmFit)
  TrainRMSE TrainRsquared    method
1  15.11114     0.7004068 svmLinear

> getTrainPerf(bayesglmFit)
  TrainRMSE TrainRsquared   method
1  14.77404     0.7115786 bayesglm

> getTrainPerf(nnetFit)
  TrainRMSE TrainRsquared method
1  14.51149     0.7235189   nnet

> getTrainPerf(rfFit70)
  TrainRMSE TrainRsquared method
1  7.835821     0.9188825     rf
```

Instance-based methods such as *k*-nearest neighbors (kNN) delivered reasonable overlay with no large systematic deviations from the $y = x$ line (*1*). However, there were often significant differences between the predicted and actual yields, indicating the model did not generate very accurate predictions. This was seen quantitatively with a test set RMSE of 16.29 and a coefficient of determination of 0.80. Algorithms based on support vector machines (SVM), Bayes generalized linear model (GLM), and linear regression were also tested (*1*). These model types all exhibited similar characteristics in their respective calibration plots. Due to the minimum experimental yield of 0% and the existence of negative predicted values for yield, these three models exhibit systematic deviation from a zero intercept in the best-fit line. Coding negative yield predictions as zeroes provides only marginal improvements for these models in quality-of-fit metrics (~0.5% decrease in RMSE and 0.02 increase in R^2 value). Neural networks were also explored as a supervised learning method. Using a single layer neural network gave similar performance to the other learning methods. The number of hidden layers in the networks was not extensively explored, and the use of deep neural networks may allow for a more predictive model (*1, 5*). Unfortunately, none of the aforementioned algorithms were able to predict reaction yield within a RMSE of 15% or an R^2 value greater than 0.82 (Fig. S22).

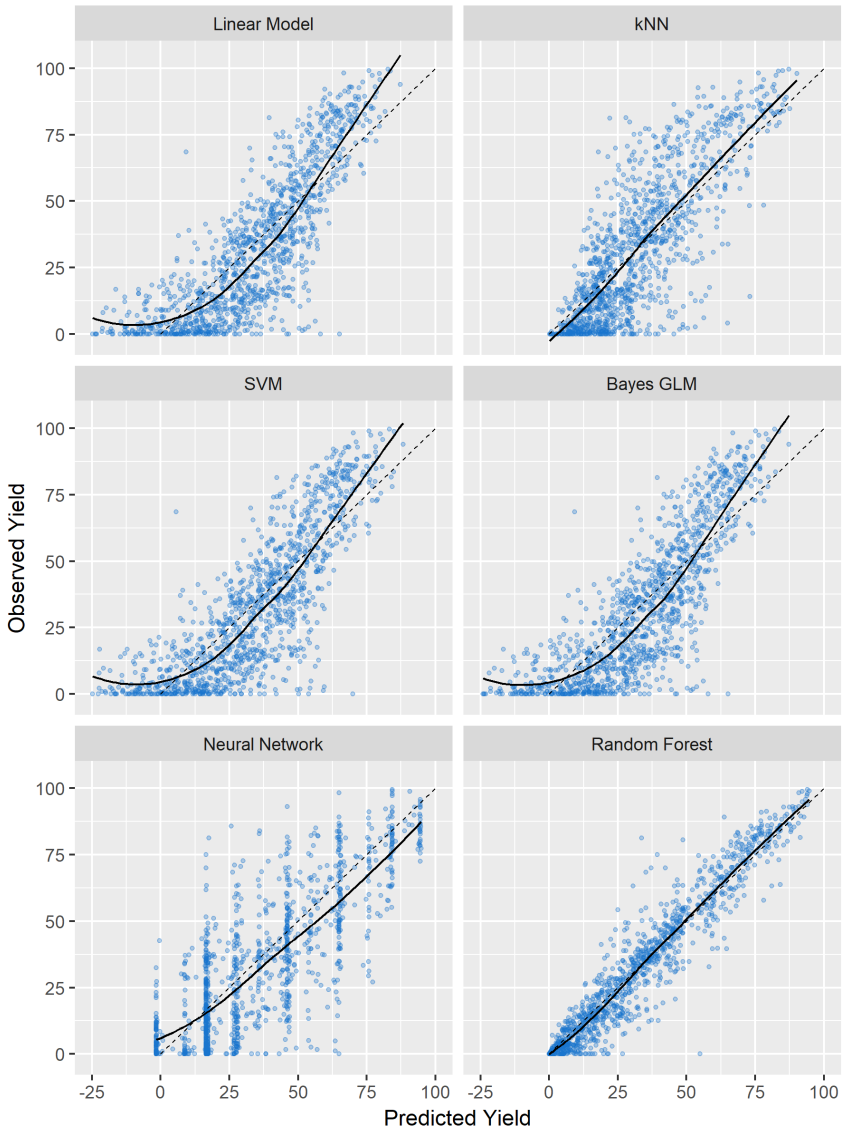


Figure S21. Training set performance plots using various machine learning models.

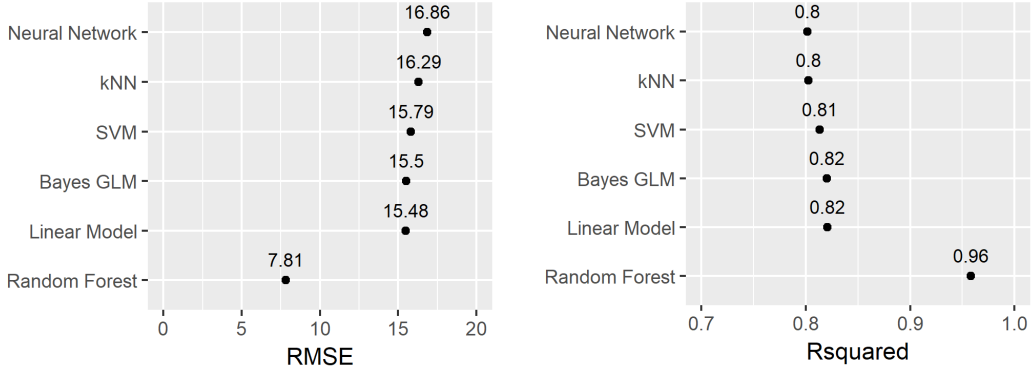


Figure S22. Test set performance using various machine learning models.

A random forest model delivered significantly improved performance in terms of predictive accuracy (*1, 4*). Random forest is an ensemble learning method for regression and classification. It works by taking many random samples from the data (with replacement) and constructing decision trees. For regression, these individual trees output predictions which can be aggregated to generate an overall prediction. The calibration plot for the random forest model demonstrates high fidelity between the predicted and observed yields. Furthermore, the root mean square error for the random forest model in predicting the test set was found to be 7.81, a very low number considering the combined experimental and analytical error is likely at least 5%. Finally, the test set predicted vs. observed plot has a very high coefficient of determination of 0.96. Clearly, the random forest algorithm provides significant improvements over other machine learning methods in terms of predictive accuracy.

The relative importance of descriptors used to construct this random forest model was also explored as a possible method for probing reaction mechanism. The most robust and informative measure of a variable's importance is the percent increase in mean square error of a model when values for that variable are randomly shuffled and the model is retrained. This percentage is measured relative to the initial model built with the unpermuted variable, and should not be confused with the RMSE metric shown elsewhere. As seen in Fig. S23, the most important variable in predicting reaction outcomes is the calculated NMR shift for *C3 on the additive. The second and fourth most predictive descriptors are the E_{LUMO} and *O1 electrostatic charge of the additive. Together, these results suggest that the additive acting as an electrophile strongly influences reaction outcomes. Such a finding is consistent with oxidative addition of the N–O bond being a source of deleterious side reactivity. Interestingly, there is no obvious relationship between the *C3 NMR shift in isolation and reaction yield (Fig. S24, Top). Using the top five most influential descriptors from the trained random forest model also does not produce a predictive model (Fig. S24, Bottom).

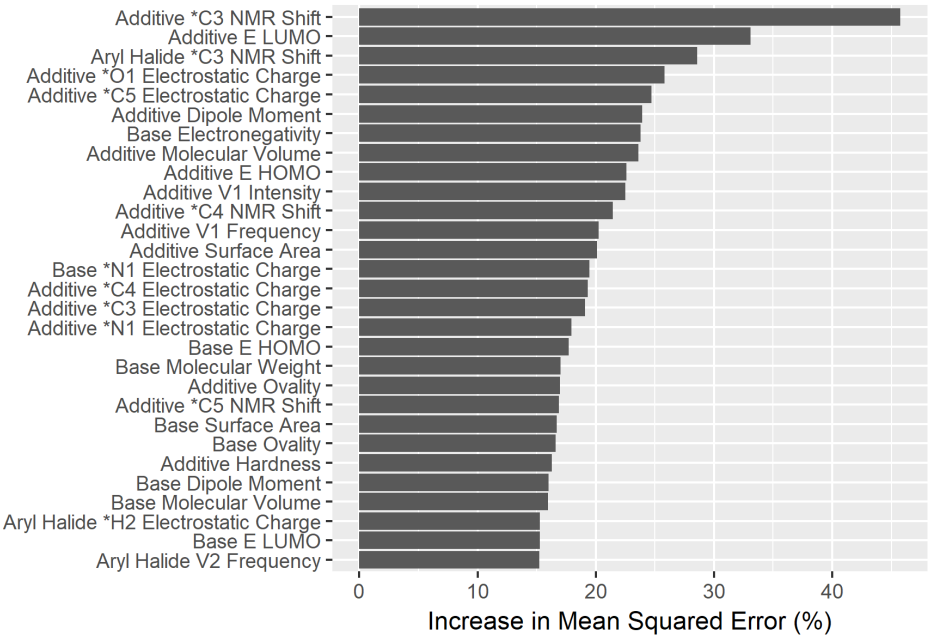


Figure S23. Descriptor importance of trained random forest model.

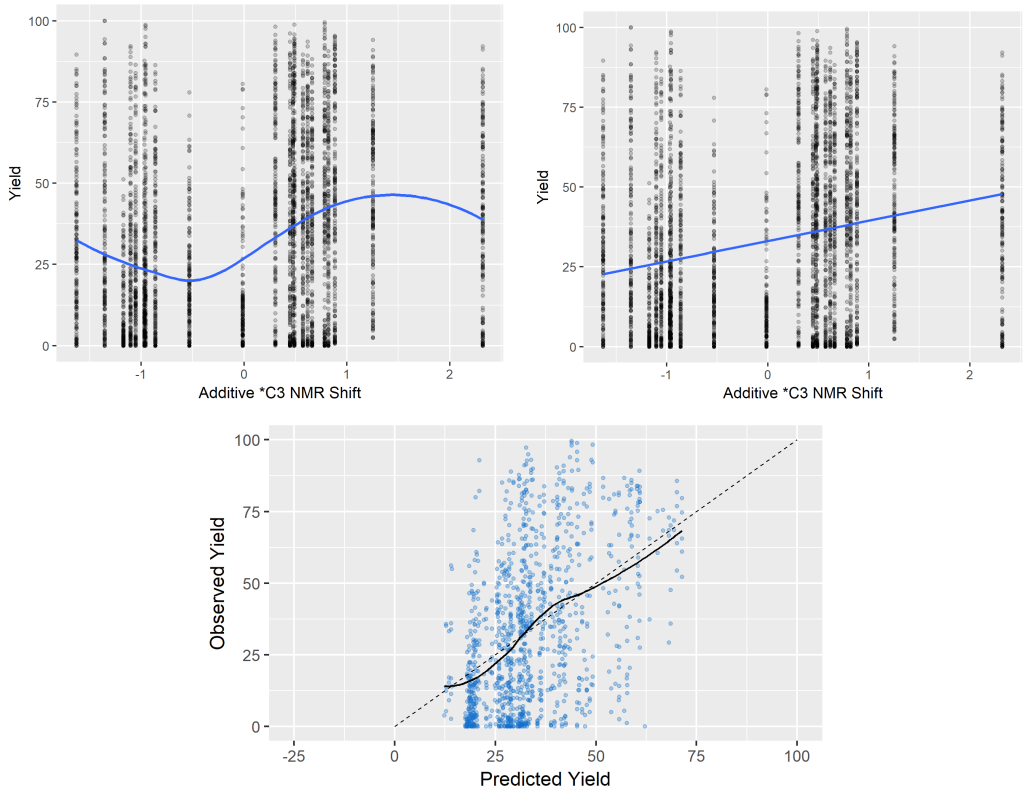


Figure S24. (Top) Plots of yield vs. *C3 NMR shift modeled using Loess and linear curves (R^2 = 0.23). (Bottom) Linear Model calibration plot using top 5 RF descriptors (R^2 = 0.45, RMSE = 24.1).

We were also interested in whether a random forest model would allow for effective prediction using a sparse reaction matrix. The ability to perform accurate prediction under sparsity would effectively increase the reaction space that can be explored using the same number of experiments. For instance, if a reaction matrix containing half as many points were nearly as predictive, an experimentalist could halve the number of experiments and maintain similar predictive power. Alternatively, one could run the same number of experiments, but distributed over double the reaction space. Both of these strategies have the potential to accelerate the optimization of chemical reactions or scope elucidation. We were very pleased to discover that enhanced predictive power over other methods could be achieved using a smaller subset of the training data. A gradual erosion in predictive accuracy is witnessed from 70% of the data (the entire training set) down to 2.5% of the full dataset (Fig. S25 and Fig. S26). Surprisingly, the random forest algorithm trained on only 5% of the reaction data outperformed any regression technique using 70% of the same reaction data.

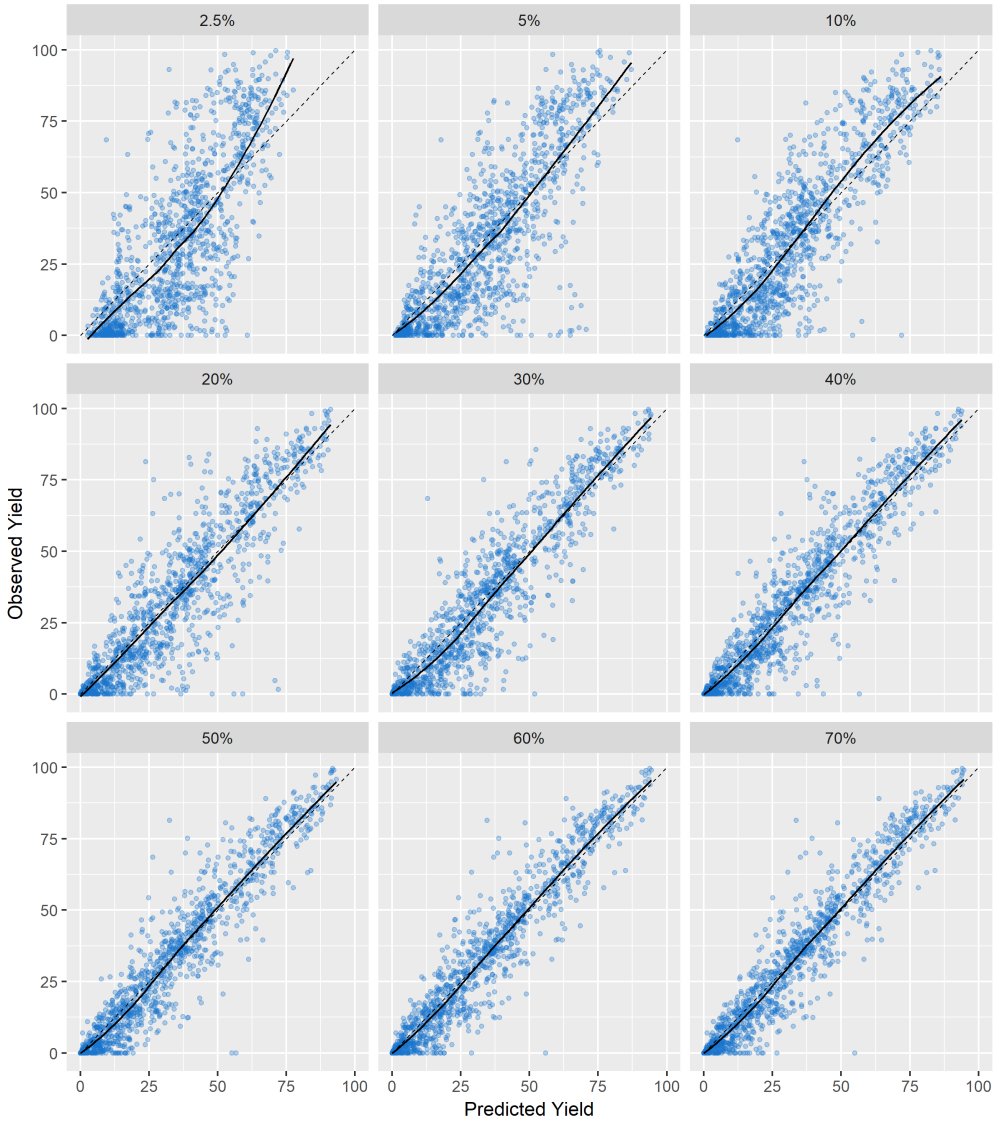


Figure S25. Calibration plots for random forest model trained using sparse data.

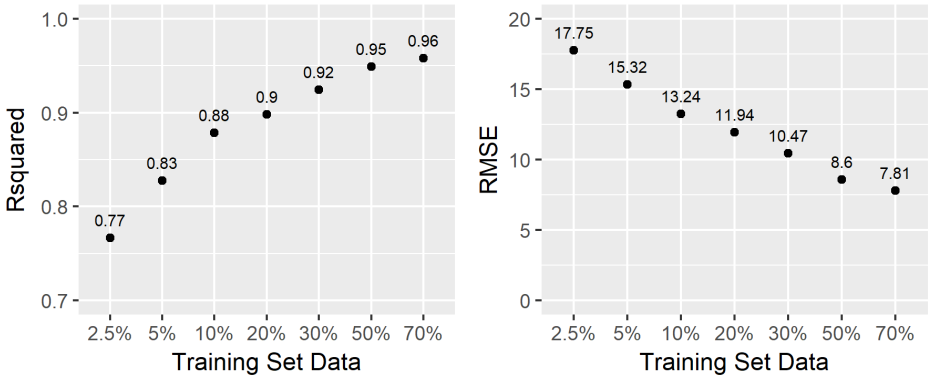


Figure S26. Test set performance of random forest model under sparsity (percentage shown based on all data – 70% is all of the training data).

We also explored the ability of the model to predict outcomes for reactions containing chemicals not included in the model's training data. Highly predictive out-of-sample performance would allow for smaller reaction matrices to be run using libraries for each reagent class. If these experimental reaction yields could subsequently be used to predict reaction outcomes for different compounds, that would effectively decrease the number of reactions that would need to be run. Such a technique would be especially useful in medicinal chemistry, where the optimal conditions could be determined without needing to expend precious material. Furthermore, since the decision to use calculated descriptors obviates the need to obtain any spectroscopic information, a substrate's performance in a coupling reaction could be predicted prior to its preparation.

For out-of-sample estimation, we evaluated whether the performance of the additives from Plate 3 could be predicted from the results of Plates 1 and 2. In this case, the results of 14 additives were used to predict the outcomes with eight distinct additives. Plates 1 and 2 were used as the training set, using leave-one-out cross validation by additive. The cross-validation performance of the model is shown below:

```
> getTrainPerf(rfFit.LOO)
  TrainRMSE TrainRsquared method
1  3.730449     0.9715724     rf
```

Intriguingly, the heterocycles in the test set contained functional groups not included in the training data, such as dibenzylamine, pyrrole, furan, and thiophene. Such an analysis produced the calibration plots in Fig. S27. On average, the out-of-sample RMSE was 11.0% and the R^2 value of 0.912. Surprisingly, none of the additives experienced significant systematic deviations from what was predicted by the model. The high predictive ability of the model suggests that the effects of these substituents on reaction outcome were well-captured by the descriptors and that out-of-sample estimation using a random forest model is effective.

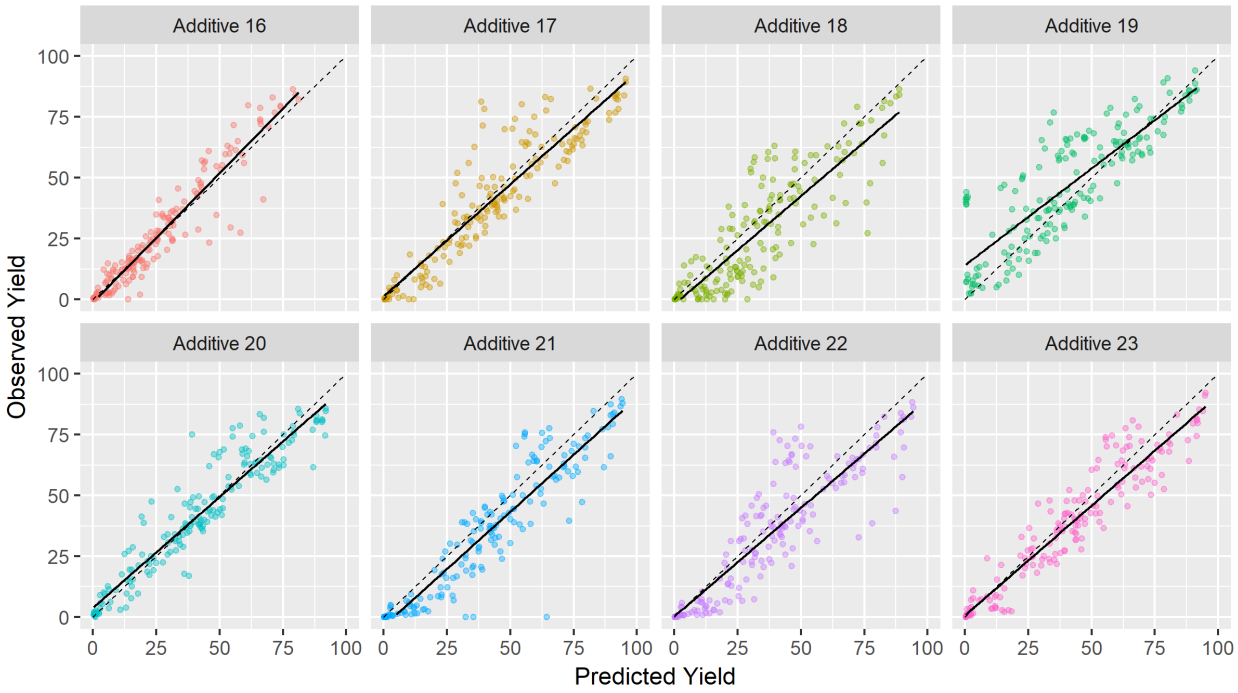


Figure S27. Out-of-sample performance of random forest model.

Random forest was also used to model each aryl halide individually. The full dataset was first split between aryl chlorides, bromides, and iodides. Each of these subgroups was then split into training and test sets using a 70/30 split. A random forest model was generated using each of these training sets and calibration plots were created using the test sets (Fig. S28). A summary of the performance metrics for these models is shown in Fig. S29. All three models demonstrated good performance and had similar RMSE and R^2 values. The lower RMSE value for aryl chlorides is likely due to the disproportionate number of yields near zero.

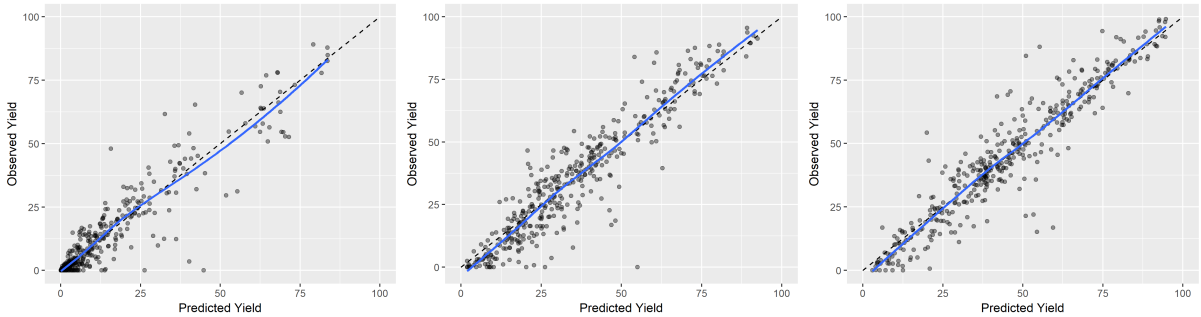


Figure S28. Random forest for aryl halides (left to right: ArCl, ArBr, ArI).

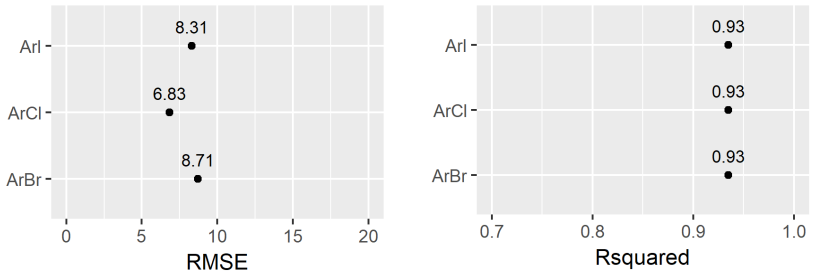


Figure S29. Test set performance of random forest model trained on individual aryl halides.

We were also interested in exploring the limits of out-of-sample prediction. First, we trained a random forest model on all of the reactions containing aryl bromides. This model was then tested on aryl chlorides and aryl iodides (Fig. S30). The aryl bromide model overpredicts the performance of aryl chlorides, meaning that these compounds are less activated than the model predicts based on extrapolating from the aryl bromide training set. When this random forest model is used to predict the performance of aryl iodides, the opposite situation is manifest. Namely, the model underpredicts the activity of aryl iodides, although this systematic error is smaller in magnitude than with the aryl chlorides.

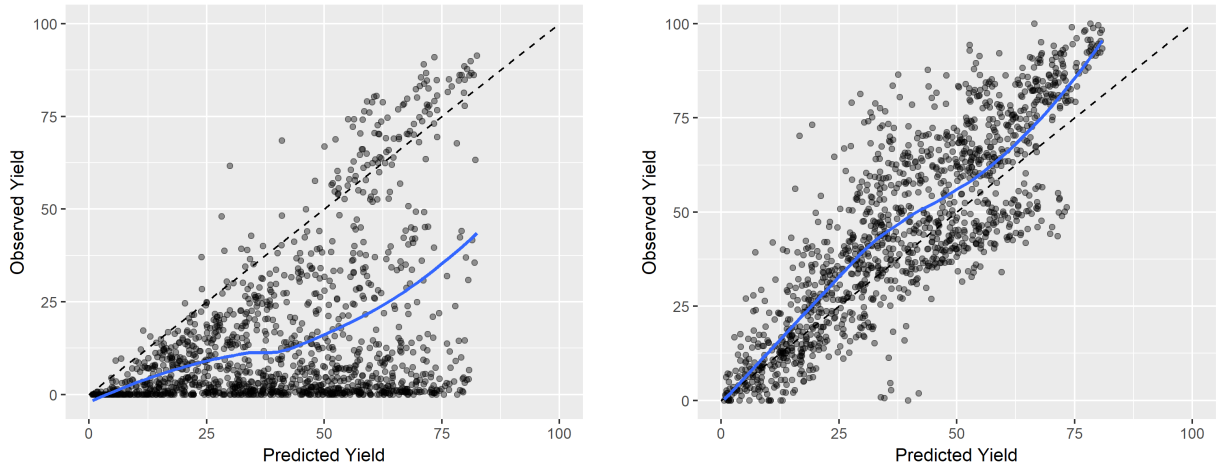


Figure S30. Random forest model trained on aryl bromides reaction and tested on aryl chlorides (left) and aryl iodides (right).

We also built a random forest model using only nonpyridyl aryl halides and tested it on the pyridyl substrates. As seen in Fig. S31, this model systematically underpredicts the yields of pyridyl aryl halides. As with the extrapolation of the aryl bromide model to aryl chlorides and iodides, the model predicts reactivity more similar to the training set than what is observed in the test set. Finally, we trained a RF model using only the yields under 80% to test the extrapolative ability of the method to higher yields. The calibration plot in Fig. S32 reveals that this strategy was ineffective using our data.

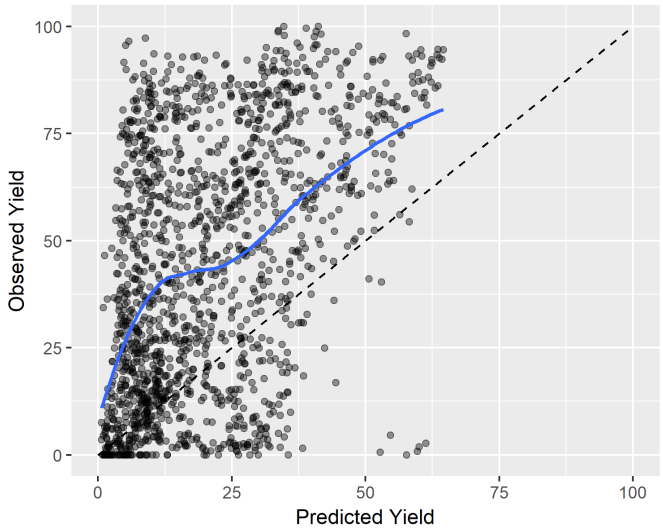


Figure S31. Random forest model trained on nonpyridyl aryl halides and tested on pyridyl aryl halides.

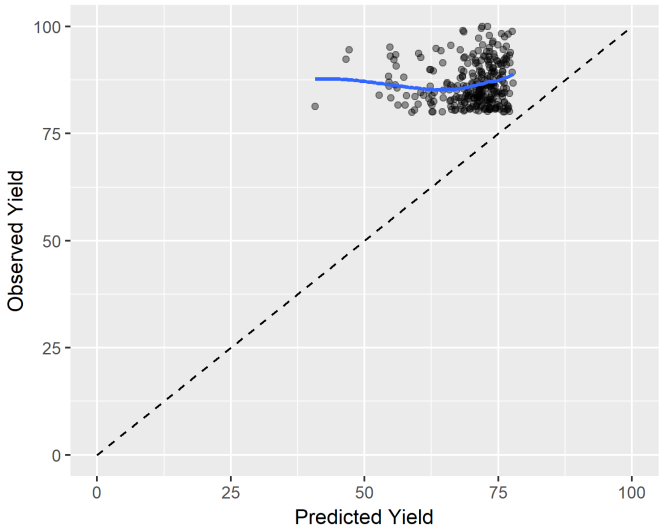


Figure S32. Random forest model trained on yields less than 80% and tested on yields greater than 80%.

Using Principal Component Analysis (PCA) with ML algorithms

A variable importance plot is generated by randomizing the values of a given descriptor followed by retraining of the model. Descriptors that are both important and unique will appear to be most influential in a variable importance plot. If a model is trained on a descriptor set that contains many highly correlated parameters, then randomizing one of these correlated parameters will not impact significantly the model's RMSE. If we wish to use variable importance plots to generate mechanistic hypotheses, then dealing with collinear descriptors would help overcome this limitation. This could be accomplished with principal component analysis. PCA reduces data sets with correlated variables by creating fewer orthogonal linear combinations of correlated descriptors.

The following lines of code were included in the analyze.R script prior to training the random forest model:

```
# perfom PCA
output.PCA <- prcomp(output, scale = T, center = T)
output.PCA$rotation # loadings
output.PCA$x # scores
summary(output.PCA) # variance/PC
output.loadings <- abs(output.PCA$rotation[,c(1:15)])

output.scores <- output.PCA$x
output.scores <- output.scores[, c(1:15)]

# load user-created yield data (label reactions w/o yield data as NA)
yield.data <- as.data.frame(read.csv("yields.csv", header=TRUE,
stringsAsFactors=FALSE))

# append the yield data to the output table
output.table <- cbind(output.scores, yield.data)
```

From the summary function we see that principal component (PC) 1 to 15 account for 95% of the variance. An output table was created from the scores of PC1 to PC15 prior to appending the yield data and training the model. The random forest model obtained with the 15 PCs has an R^2 value of 0.91 and RMSE of 9.91%, consistent with the reduced set of data. From the variable importance plot we see which PC has a greater impact on the model's RMSE. We can use the PC coefficients (loadings) to trace model analysis back to the full set of descriptors as follows.

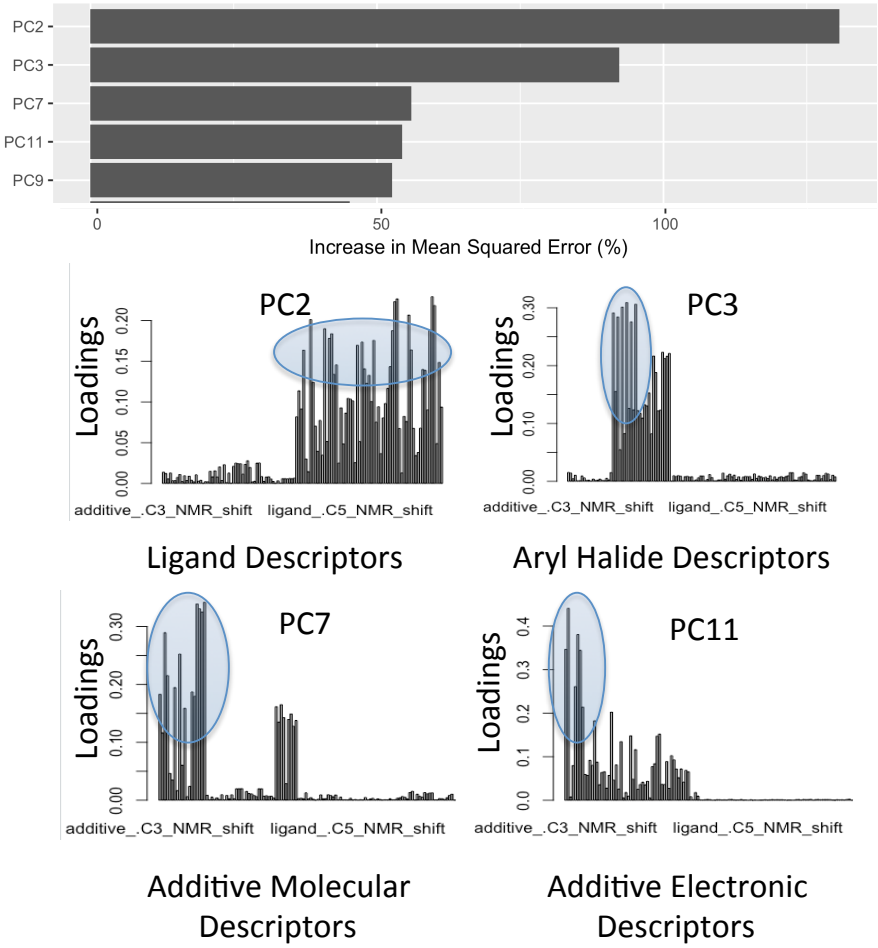


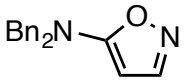
Figure S33. Constructing a RF model using PCA vectors as parameters. Top: top 5 PC vectors are shown. Below: descriptor coefficients for top 4 PC vectors.

Looking at the coefficients of PC2, primarily composed of ligand descriptors (ligand molecular vibration frequencies and intensities), it is interesting to note that it now appears to be highly influential. This is likely a result of PCA reducing the previously large number of highly correlated ligand descriptors. Next in line of important PCs is PC3, which is primarily composed of aryl halide electronic descriptors (*C1, *C2, *C3, *C4 electrostatic charges). This is followed by PC7 (additive molecular descriptors) and PC11 (additive *C3 NMR shift, additive *C3 electrostatic charge, additive *C5 NMR shift, additive *N1 electrostatic charge) as moderately important principal component vectors. While the ligands seem to play a significant role in the success of the reaction, the aryl halide and additive electronic properties appear to be important nonetheless. This is in line with what was observed from the random forest model built with the full set of ligand descriptors.

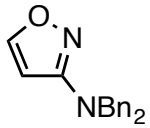
IV. Isoxazole Preparation

All aryl halides, palladium catalysts, and bases used in the Buchwald-Hartwig cross coupling are commercially (Sigma Aldrich, Fisher) available and were used without further purification. Twenty of the twenty-three isoxazole additives were also commercially (Enamine, Sigma Aldrich, Combi-Blocks) available. The remaining three isoxazoles were prepared according to literature procedures and characterized by ^{1}H-NMR, ^{13}C-NMR, HRMS (ESI-TOF), and FTIR.

Synthesis of Isoxazole Additives



***N*,*N*-dibenzylisoxazol-5-amine.** Prepared according to a known procedure (*40*).
In a 20-mL scintillation vial equipped with a magnetic stir bar was added 5-aminoisoxazole (900 mg, 10.7 mmol, 1.0 equiv), acetonitrile (42.8 mL) and sodium hydride (60% dispersed in mineral oil, 1.286 g, 32.1 mmol, 3 equiv). Benzyl bromide (3.186 mL, 26.8 mmol, 2.5 equiv) was added and the reaction stirred at room temperature for 24 h. The reaction mixture was quenched with water (50 mL) and brine (50 mL), extracted with ethyl acetate (2 x 100 mL), dried over $MgSO_4$, and concentrated *in vacuo*. The crude product was purified by column chromatography (50 g column, 2%-16% EtOAc/Hex). Fractions containing product were collected and given to Lotus Separations for further purification. Yield: 1.13 g (40%).
^{1}H NMR (500 MHz, Chloroform-*d*): δ = 7.99 (d, *J* = 3.0, 2.5 Hz, 1H), 7.36 – 7.20 (m, 10H), 4.93 (q, *J* = 2.9, 2.5 Hz, 1H), 4.48 (d, *J* = 2.2 Hz, 4H) ppm.
^{13}C NMR (126 MHz, Chloroform-*d*): δ = 170.42, 152.50, 136.46, 128.88, 127.86, 76.82, 52.51 ppm.
HRMS (ESI-TOF): Calculated for $C_{17}H_{17}N_2O$ ($[M+H]^+$) 265.1335, found 265.1332.
FTIR (thin film): 2989, 2868, 1382, 1135 cm^{-1}.



***N*,*N*-dibenzylisoxazol-3-amine.** Prepared according to the procedure above.
In a 25-mL round-bottom flask equipped with a magnetic stir bar and condenser was added 3-aminoisoxazole (350 μL, 5.0 mmol, 1.0 equiv), acetonitrile (5 mL) and potassium carbonate (2.07 g, 15 mmol, 3 equiv). Benzyl bromide (1.49 mL, 12.5 mmol, 2.5 equiv) was added and the reaction stirred at 85 °C for 24 h. The reaction mixture was quenched with water (50 mL) and brine (50 mL), extracted with ethyl acetate (2 x 100 mL), dried over $MgSO_4$, and concentrated *in*

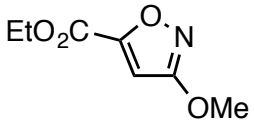
vacuo. The crude product was purified by column chromatography (50 g column, 2%-16% EtOAc/Hex). Yield: 546 mg (21%).

^{1}H NMR (500 MHz, Chloroform-*d*): δ = 8.09 (d, *J* = 1.8 Hz, 1H), 7.38 – 7.22 (m, 10H), 5.91 (d, *J* = 1.8 Hz, 1H), 4.50 (s, 4H) ppm.

^{13}C NMR (126 MHz, Chloroform-*d*): δ = 166.40, 158.68, 137.57, 128.75, 127.74, 127.50, 95.43, 52.70 ppm.

HRMS (ESI-TOF): Calculated for $C_{17}H_{17}N_2O$ ($[M+H]^+$) 265.1335, found 265.1332.

FTIR (thin film): 2919, 1586, 1512, 1451, 690 cm^{-1}.



Ethyl-3-methoxyisoxazole-5-carboxylate. Prepared according to a known procedure (*41*).

In a 50-mL round-bottom flask equipped with a magnetic stir bar and condenser was added 3-methoxyisoxazole-5-carboxylic acid (1.0 g, 7.0 mmol, 1.0 equiv), anhydrous ethanol (20 mL) and hydrogen chloride solution in diethyl ether (2.0 M, 10 mL). The mixture was refluxed for 2 h. After cooling to room temperature, the reaction mixture was neutralized with $NaHCO_3$, diluted with DCM (50 mL). The aqueous layer was extracted with DCM (3 x 50 mL), the organic layers collected, dried over $MgSO_4$, filtered and reduced. The crude product was purified by column chromatography (100 g column, 5%-20% EtOAc/Hex). Yield: 1.09 g (84%).

^{1}H NMR (500 MHz, Chloroform-*d*): δ = 6.52 (s, 1H), 4.40 (q, *J* = 7.1 Hz, 2H), 4.02 (s, 3H), 1.39 (t, *J* = 7.1 Hz, 3H) ppm.

^{13}C NMR (126 MHz, Chloroform-*d*): δ = 172.17, 160.81, 156.82, 100.61, 62.40, 57.63, 14.25 ppm.

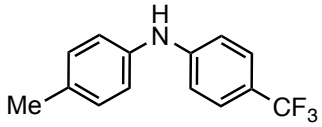
HRMS (ESI-TOF): Calculated for $C_7H_{10}NO_4$ ($[M+H]^+$) 172.0604, found 172.0602.

FTIR (thin film): 2983, 1739, 1512, 1414, 1302, 1284, 1098 cm^{-1}.

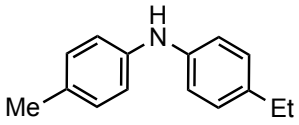
V. Products

Reaction products were isolated from a large-scale reaction. Products were characterized (see below) and analyzed by UPLC in order to obtain retention times (R_f) as shown in Table S3.

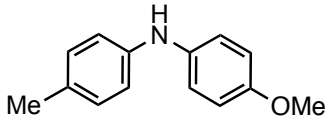
General Procedure: An oven-dried glass tube was charged with aryl bromide (1 mmol), toluidine (1.5 equiv), tBuOK (1.5 equiv), $Pd_2(dba)_3$ (0.02 equiv), BrettPhos (0.08 equiv) and toluene (5 mL). N_2 gas was bubbled through the reaction mixture for 15 minutes, after which it was heated at 80 °C overnight. After cooling to room temperature, the reaction mixture was diluted with EtOAc (10 mL), passed through a short plug of silica and reduced under pressure. The crude material was purified by column chromatography (*42*).



4-methyl-*N*-(4-(trifluoromethyl)phenyl)aniline. ^{1}H- and ^{13}C-NMR chemical shifts matched literature values (*42*).
^{1}H NMR (500 MHz, Chloroform-*d*): δ = 7.44 (d, *J* = 8.5 Hz, 2H), 7.15 (d, *J* = 8.1 Hz, 2H), 7.07 (d, *J* = 8.5 Hz, 2H), 6.98 (d, *J* = 8.5 Hz, 2H), 5.94 (s, 1H), 2.34 (s, 3H) ppm.
^{13}C NMR (126 MHz, Chloroform-*d*): δ = 147.65, 138.47, 133.04, 130.17, 126.73 (q, $^3J_{C\text{-}F}$ = 3.9 Hz), 124.84 (q, $^1J_{C\text{-}F}$ = 269.6 Hz), 121.13, 121.03 (q, $^2J_{C\text{-}F}$ = 33.2 Hz), 114.71, 20.94 ppm.
^{19}F NMR (470 MHz, Chloroform-*d*): δ = –61.3 ppm.

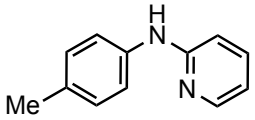


4-ethyl-*N*-(*p*-tolyl)aniline. Authentic product was synthesized according to a known procedure (*43*), and used to compare with the reaction product.
^{1}H NMR (500 MHz, Chloroform-*d*): δ = 7.14 – 7.05 (m, 4H), 7.02 – 6.95 (m, 4H), 5.55 (s, 1H), 2.61 (q, *J* = 7.6 Hz, 2H), 2.31 (s, 3H), 1.24 (t, *J* = 7.6 Hz, 3H) ppm.
^{13}C NMR (126 MHz, Chloroform-*d*): δ = 141.46, 141.14, 136.75, 130.28, 129.92, 128.73, 118.08, 117.87, 28.24, 20.77, 15.97 ppm.
HRMS (ESI-TOF): Calculated for $C_{15}H_{18}N$ ($[M+H]^+$) 212.1434, found 212.1428.
FTIR (thin film): 3414, 2914, 1606, 1509, 1457, 1302, 1247, 1176, 808 cm^{-1}.

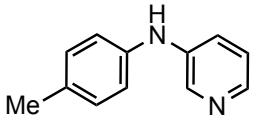


4-methoxy-*N*-(*p*-tolyl)aniline. ^{1}H- and ^{13}C-NMR chemical shifts matched literature values (*42*).
^{1}H NMR (500 MHz, Chloroform-*d*): δ = 7.05 (d, *J* = 8.0 Hz, 4H), 6.95 – 6.76 (m, 4H), 6.25 – 4.85 (m, 1H), 3.80 (s, 3H), 2.28 (s, 3H) ppm.

^{13}C NMR (126 MHz, Chloroform-*d*): δ = 154.92, 142.36, 136.60, 129.91, 129.52, 121.24, 116.72, 114.75, 55.69, 20.69 ppm.



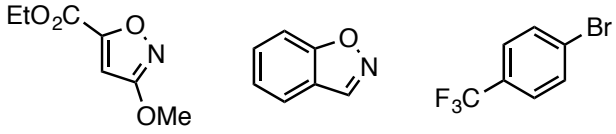
***N*-(*p*-tolyl)pyridin-2-amine.** ^{13}C-NMR shifts matched literature values (*42*). ^{1}H-NMR integrations agree with the structure above, although the multiplicity and chemical shifts for some peaks are different compared to the literature values. This may be a result of salty samples.
^{1}H NMR (500 MHz, Chloroform-*d*): δ = 8.17 (d, *J* = 4.6 Hz, 1H), 7.47 (t, *J* = 7.8 Hz, 1H), 7.20 (d, *J* = 8.3 Hz, 2H), 7.15 (d, *J* = 8.1 Hz, 2H), 6.82 (d, *J* = 8.4 Hz, 1H), 6.74 – 6.67 (m, 1H), 6.66 – 6.58 (m, 1H), 2.33 (s, 3H) ppm.
^{13}C NMR (126 MHz, Chloroform-*d*): δ = 156.55, 148.27, 137.91, 137.70, 133.04, 130.00, 121.40, 114.75, 107.90, 20.98 ppm.



***N*-(*p*-tolyl)pyridin-3-amine.** ^{1}H- and ^{13}C-NMR chemical shifts matched literature values (*42*).
^{1}H NMR (500 MHz, Chloroform-*d*): δ = 8.34 (s, 1H), 8.11 (d, *J* = 4.1 Hz, 1H), 7.34 (d, *J* = 9.4 Hz, 1H), 7.19 – 7.08 (m, 3H), 7.01 (d, *J* = 8.2 Hz, 2H), 5.72 (s, 1H), 2.32 (s, 3H) ppm.
^{13}C NMR (126 MHz, Chloroform-*d*): δ = 141.22, 140.73, 139.32, 139.10, 132.26, 130.21, 123.85, 122.59, 119.54, 20.89 ppm.

VI. Organometallic Results

A tetrakis(triphenylphosphine)palladium(0) stock solution was prepared in a 1-dram vial by dissolving $Pd(PPh_3)_4$ (138.7 mg, 0.12 mmol) in d_6-benzene (2.4 mL, 0.05 M). Similar stock solutions were made for compounds **1a** (34.8 mg, 0.203 mmol) in d_6-benzene (452 μL, 0.45 M); **1b** (37.6 mg, 0.315 mmol) in d_6-benzene (703 μL, 0.45 M); and **1c** (80.0 mg, 0.356 mmol) in d_6-benzene (790 μL, 0.45 M). All stock solutions were prepared inside a glove box. Reactions with $Pd(PPh_3)_4$ and the competition experiments were set up in 1-dram vials equipped with a magnetic stir bar according to the following table:



1a **1b** **1c**

Reaction	**1a** stock (μL)	**1b** stock (μL)	**1c** stock (μL)	C_6D_6 (μL)	Pd stock (μL)
1a	100	0	0	100	400
1b	0	100	0	100	400
1c	0	0	100	100	400
1a + **1c**	100	0	100	0	400
1b + **1c**	0	100	100	0	400

Final concentration for all reaction solutions were $Pd(PPh_3)_4$ (0.033M, 1 equiv), **1a** (0.075 M, 2.3 equiv), **1b** (0.075 M, 2.3 equiv), and **1c** (0.075 M, 2.3 equiv). The vials were stirred for 1 hour inside the glove box at room temperature and then transferred to an NMR tube for ^{31}P-NMR analysis.

Spectroscopic Characterization of 2b

Inside a glove box, a 1-dram vial was charged with isoxazole **1b** (73 mg, 0.61 mmol) and C_6D_6 (1.5 mL). A separate 1-dram vial equipped with a magnetic stirrer was charged with $Pd(PPh_3)_4$ (50 mg, 0.04 mmol), C_6D_6 (500 μL) and the **1b** stock solution (500 μL). The vial was covered with a septum cap and sealed with electrical tape. The vial was brought outside the box, and heated at 60 °C for 9 hours. After cooling to room temperature, 600 μL of the reaction mixture was transferred to an NMR tube for spectroscopic analysis (^{31}P-NMR, ^{1}H-NMR, quant^{13}C-NMR, C-APT, HSQC). The reaction mixture was also analyzed by High Resolution Mass Spectrometry (ESI-TOF), showing the presence of $Pd(PPh_3)_2$(**1b**), calculated for $C_{43}H_{36}NOP_2Pd$ ($[M+H]^+$) 750.1301, found 750.1325.

Analysis of the reaction mixture showed that in the presence of excess isoxazole, about 80% of the starting material had isomerized to 2-hydroxybenzonitrile (**3b**). This isomerization appears to originate from oxidative adduct **2b** since heating a sample of isoxazole **1b** in the absence of palladium gave no isomerized product. A proposed mechanism for the generation of **3b** from **2b** is shown below.

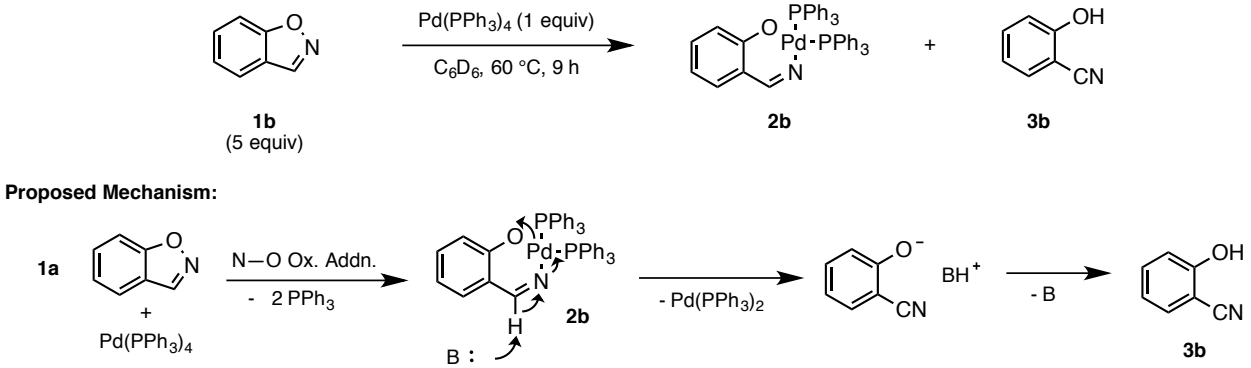
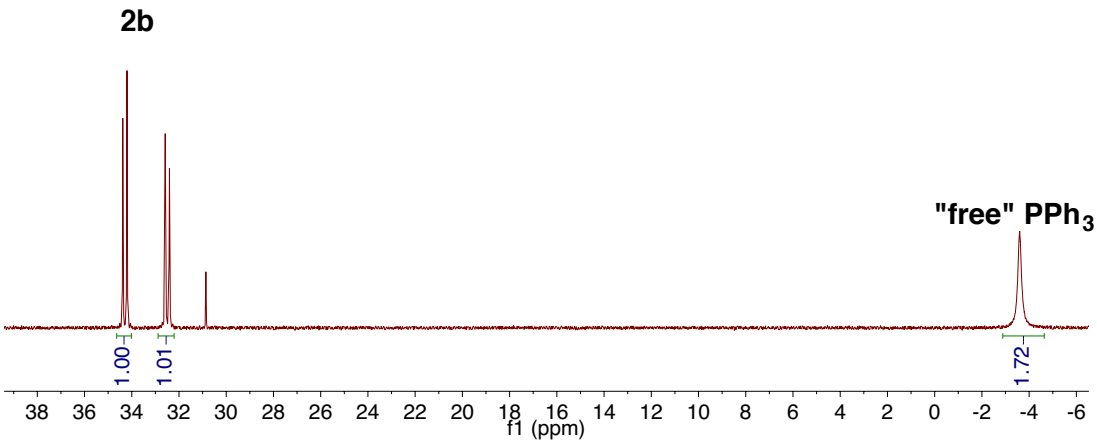


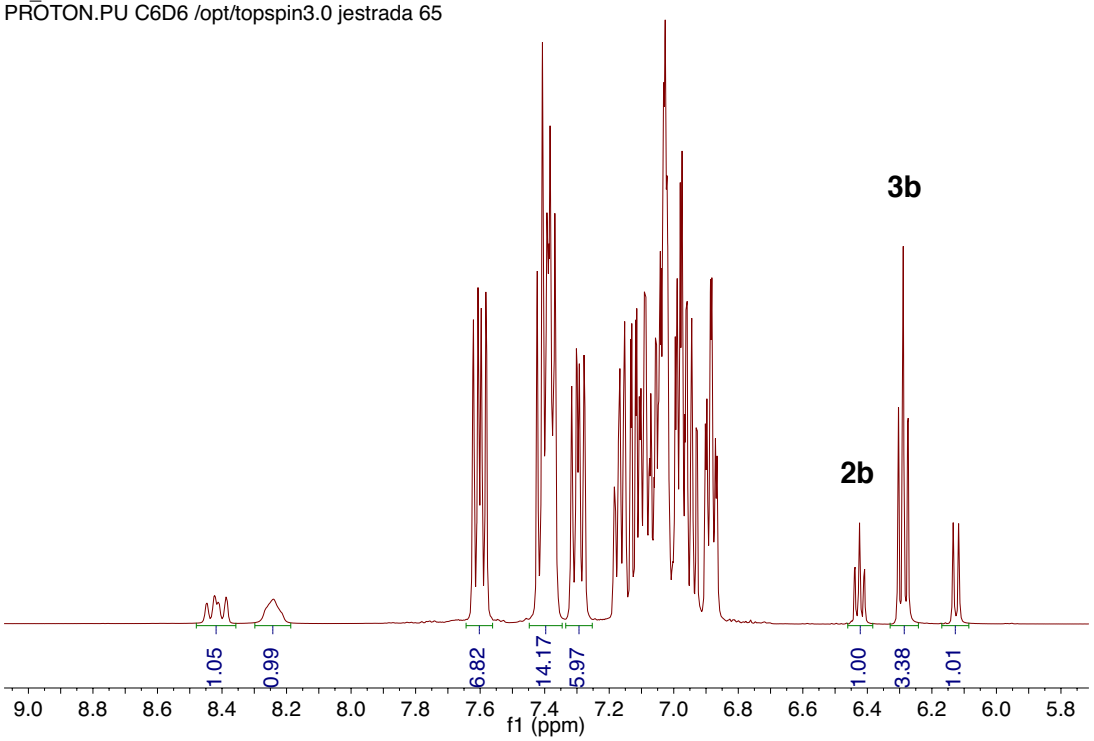
Figure S34. Top: Reaction of isoxazole **1b** with $Pd(PPh_3)_4$. Below: Proposed mechanism for formation of side product **3b** from **2b**.

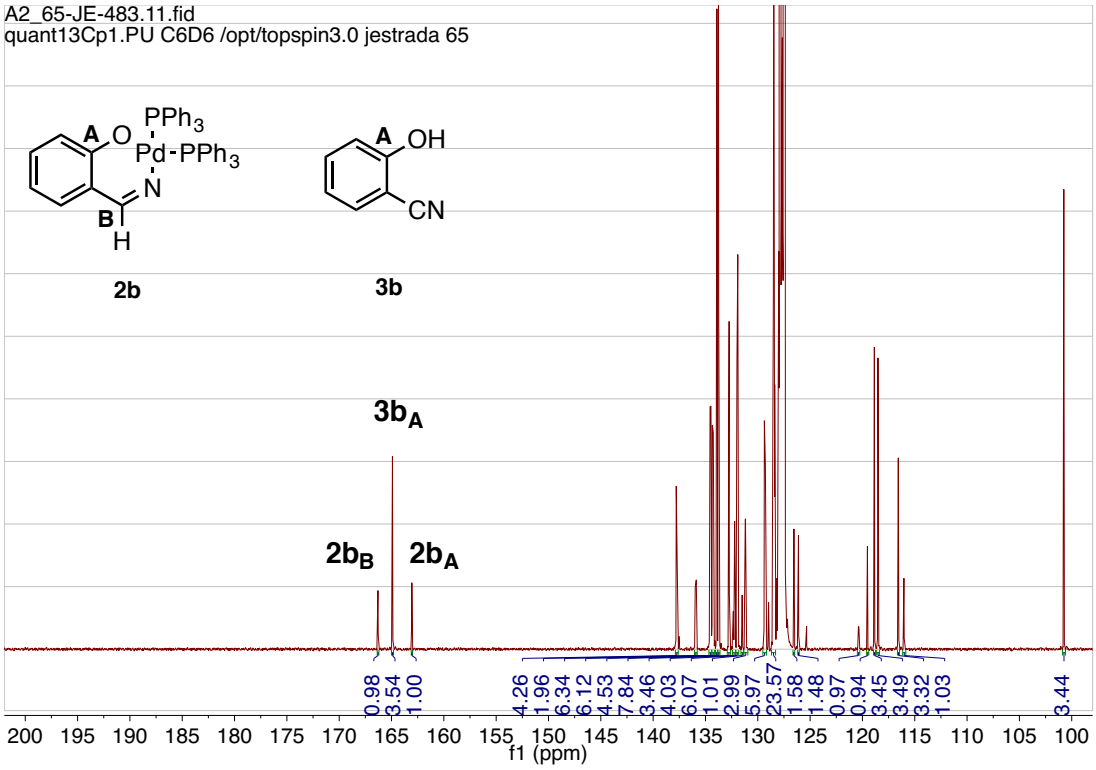
The ^{1}H-NMR spectrum of the reaction mixture shows a 1:3.4 ratio of oxidative adduct **2b** to **3b**. We proceeded to use ^{13}C-NMR for further spectroscopic evidence of the structure of oxidative adduct **2b**. The quantitative ^{13}C-NMR spectrum shows three peaks in the downfield region of 150-200 ppm with integrations matching the products ratio observed in the ^{1}H-NMR spectrum (1:3.5). In this region we expect to find the imine and phenolic carbons. The three peaks have been labeled as $\mathbf{2b_A}$, $\mathbf{2b_B}$ and $\mathbf{3b_A}$ (*vide infra*). If the structure of **2b** is as assigned, then the ^{13}C-NMR with the Attached Proton Test (C-APT) spectrum would have imine peak $\mathbf{2b_B}$ (C—H) on the opposite side of peaks $\mathbf{2b_A}$ (quaternary C) and $\mathbf{3b_A}$ (quaternary C). However, if oxidative addition had occurred at the imine C—H bond, then we would expect peak $\mathbf{2b_B}$ to be on the same side of peaks $\mathbf{2b_A}$ and $\mathbf{3b_A}$, since the carbon of $\mathbf{2b_B}$ would no longer be attached to the aldimine proton. The C-APT spectrum confirms the presence of the aldimine group in oxidative adduct **2b** consistent with N–O oxidative addition. Heteronuclear Single Quantum Correlation (HSQC) spectroscopy couples that aldimine carbon to a proton with a chemical shift of 8.4 ppm.

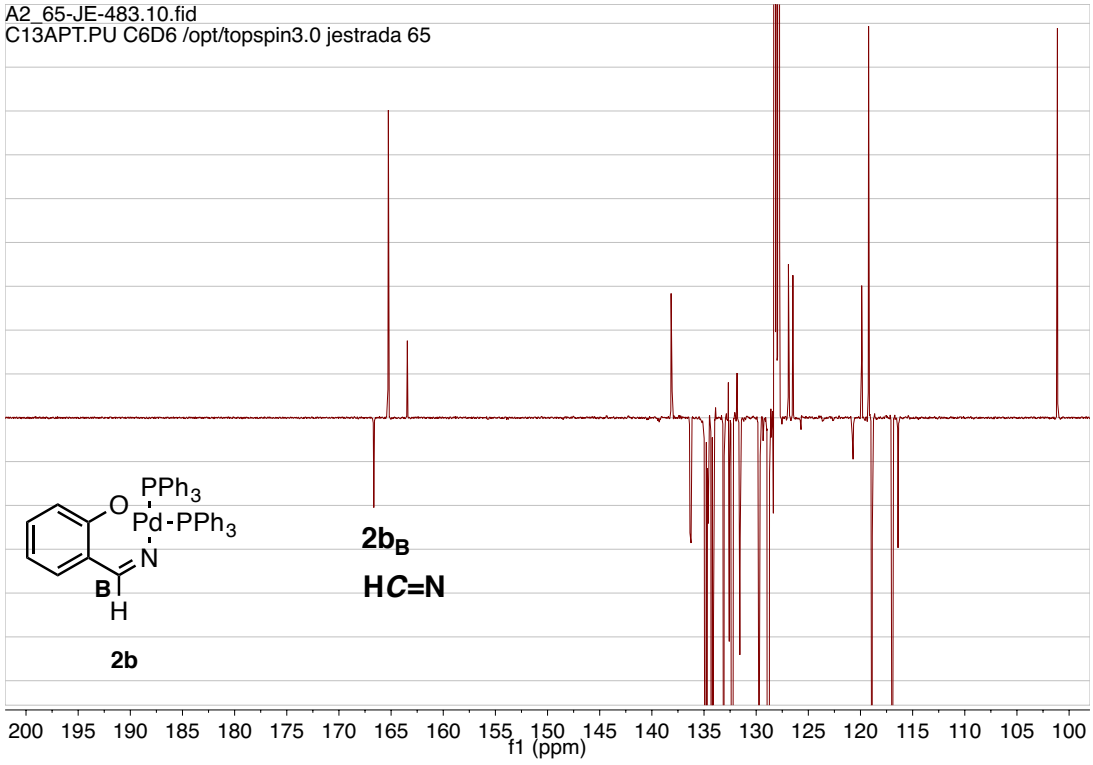
A500_JE-483.10.fid
JE-483 5eq 1b 9 h 60C in C6D6

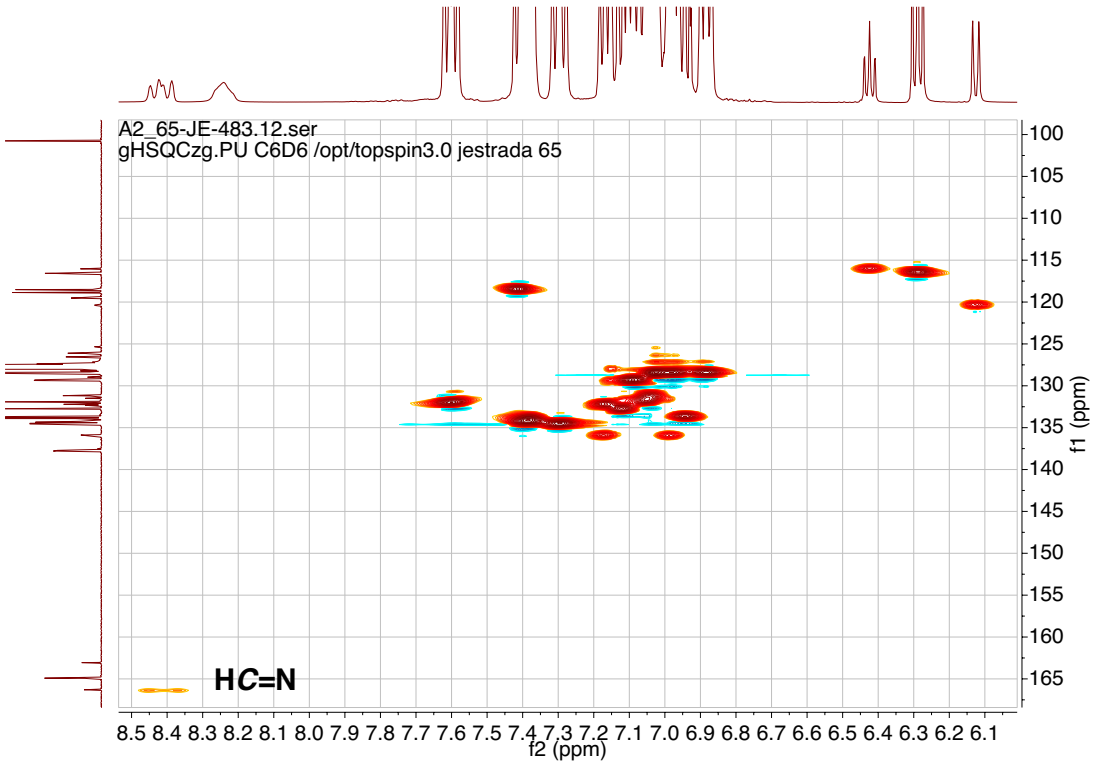


A2_65-JE-483.13.fid
PROTON.PU C6D6 /opt/topspin3.0 jestrada 65

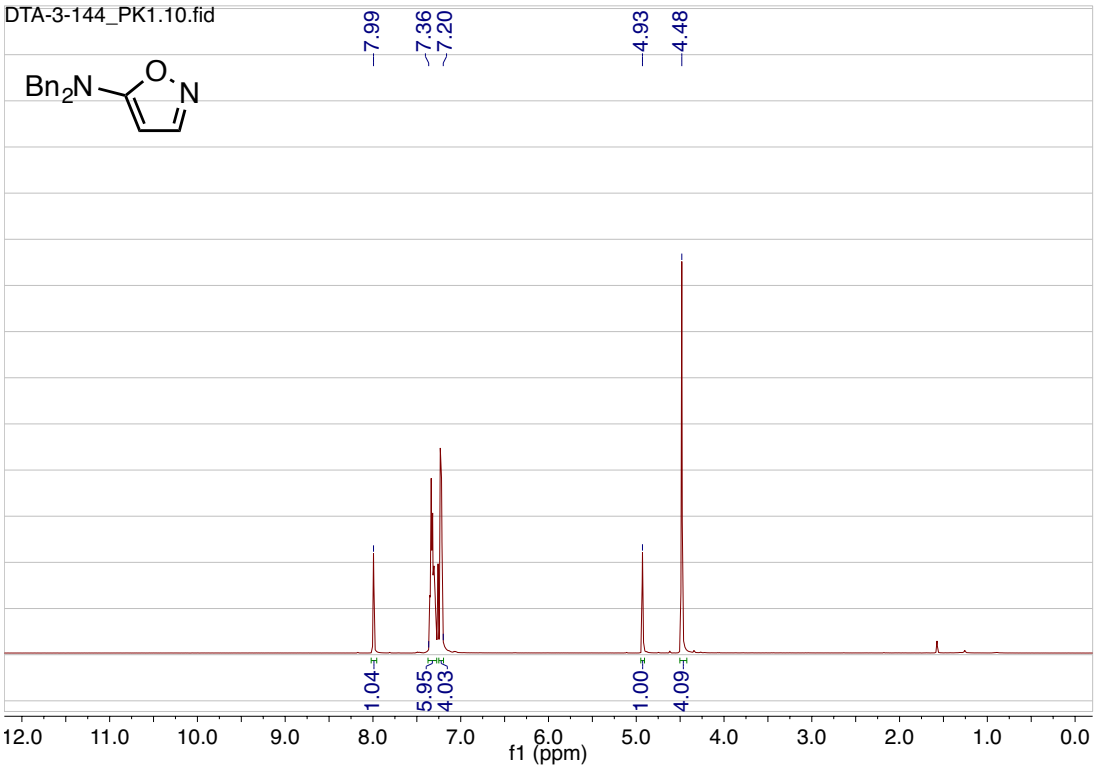


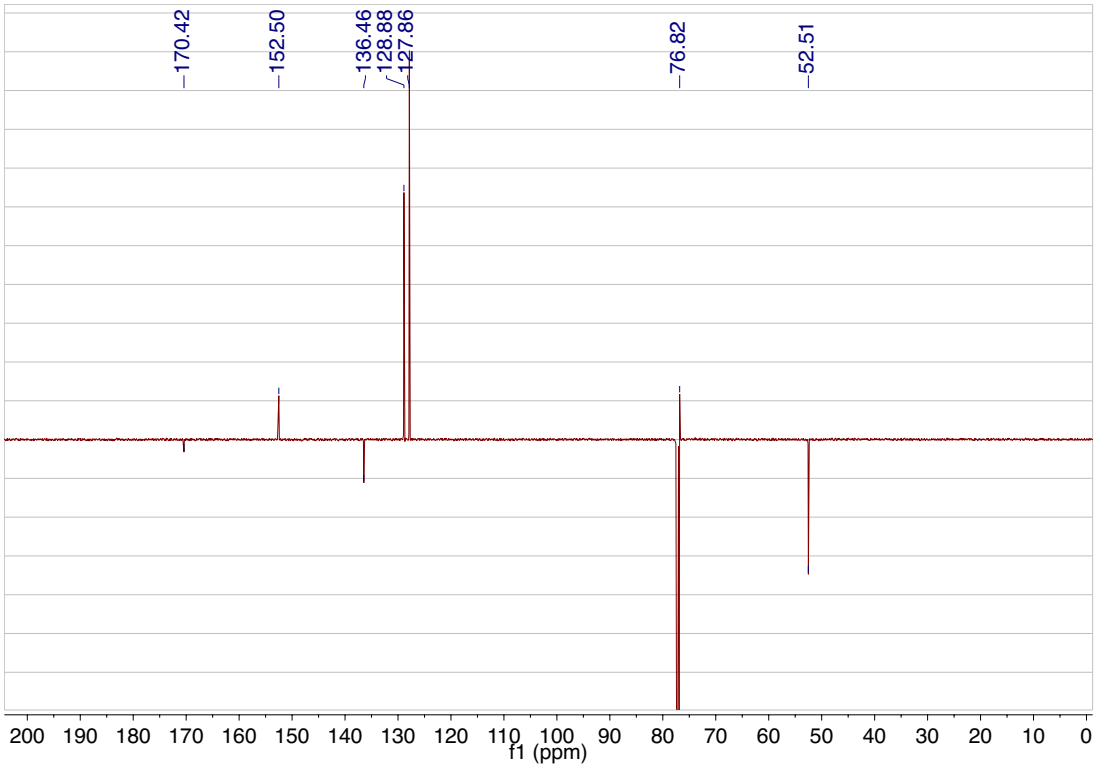

A2_65-JE-483.11.fid
quant13Cp1.PU C6D6 /opt/topspin3.0 jestrada 65
PPh3
A
O
Pd-PPh3
N
B
H
2b
A
OH
CN
3b
3bA
2bB
2bA
0.98
3.54
1.00
4.26
1.96
6.34
6.12
4.53
7.84
3.46
4.03
6.07
1.01
2.99
5.97
23.57
1.58
1.48
0.97
0.94
3.45
3.49
3.32
1.03
3.44
200 195 190 185 180 175 170 165 160 155 150 145 140 135 130 125 120 115 110 105 100
f1 (ppm)

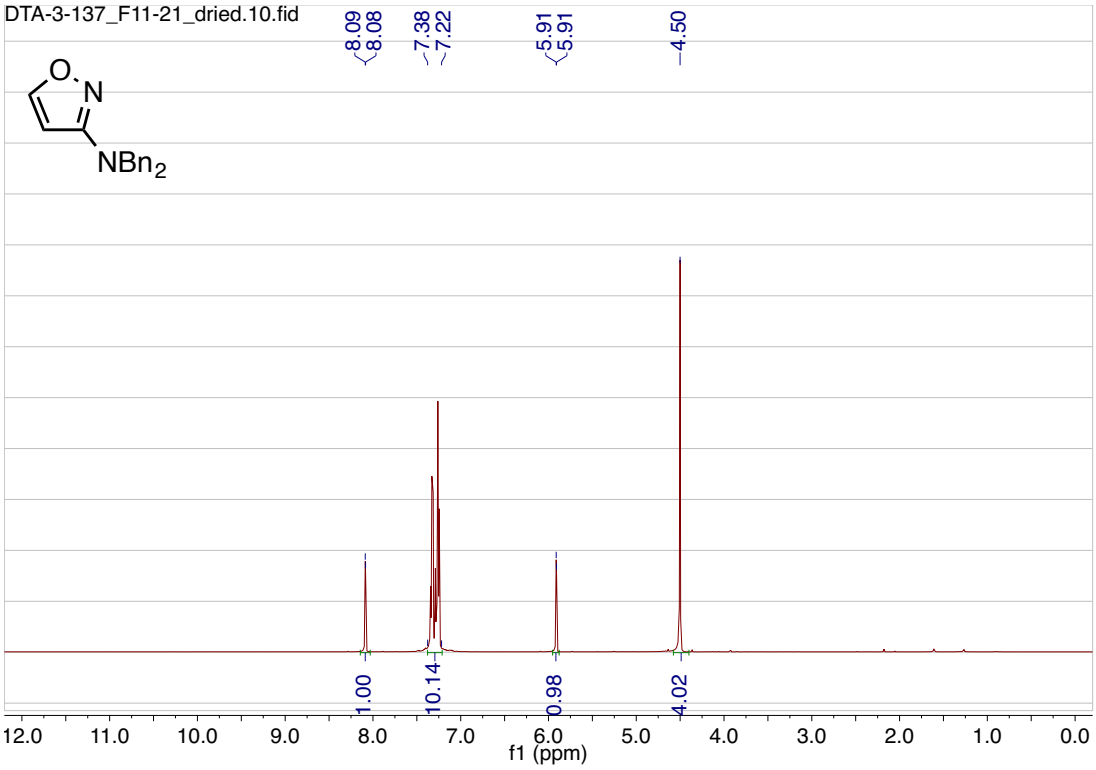

A2_65-JE-483.10.fid
C13APT.PU C6D6 /opt/topspin3.0 jestrada 65
PPh3
O
Pd-PPh3
N
B
H
2b
2bB
HC=N
200 195 190 185 180 175 170 165 160 155 150 145 140 135 130 125 120 115 110 105 100
f1 (ppm)

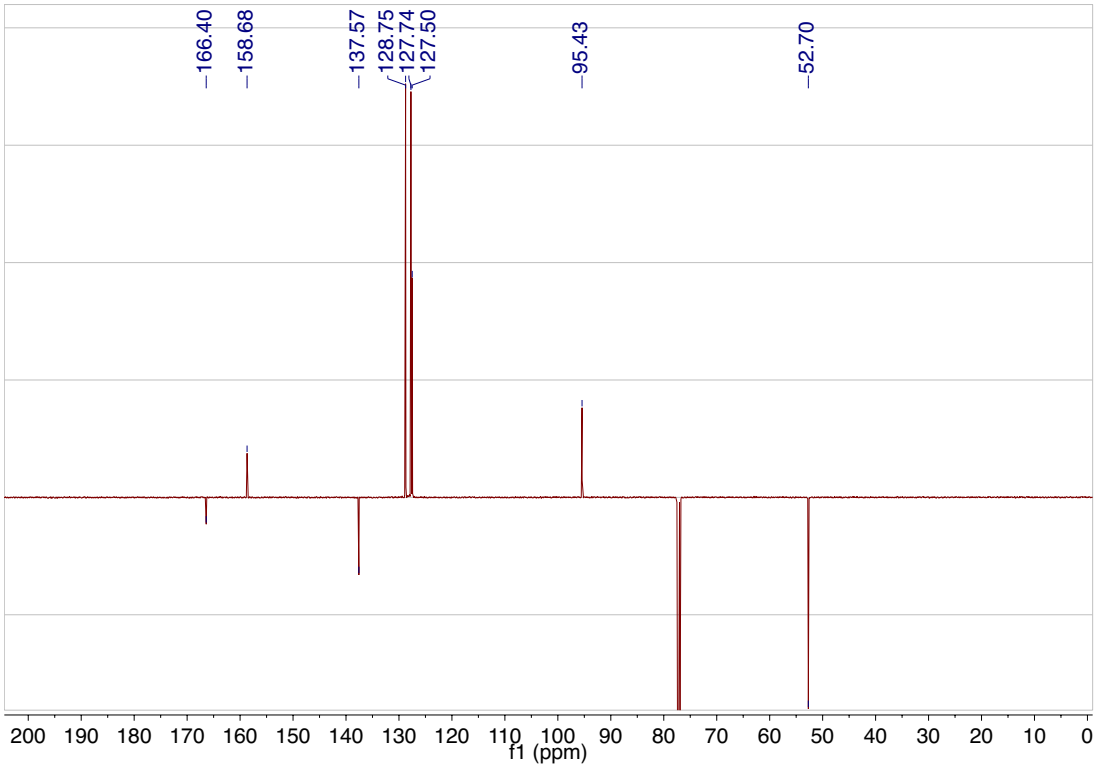

A2_65-JE-483.12.ser
gHSQCzg.PU C6D6 /opt/topspin3.0 jestrada 65
HC=N
f1 (ppm)
f2 (ppm)

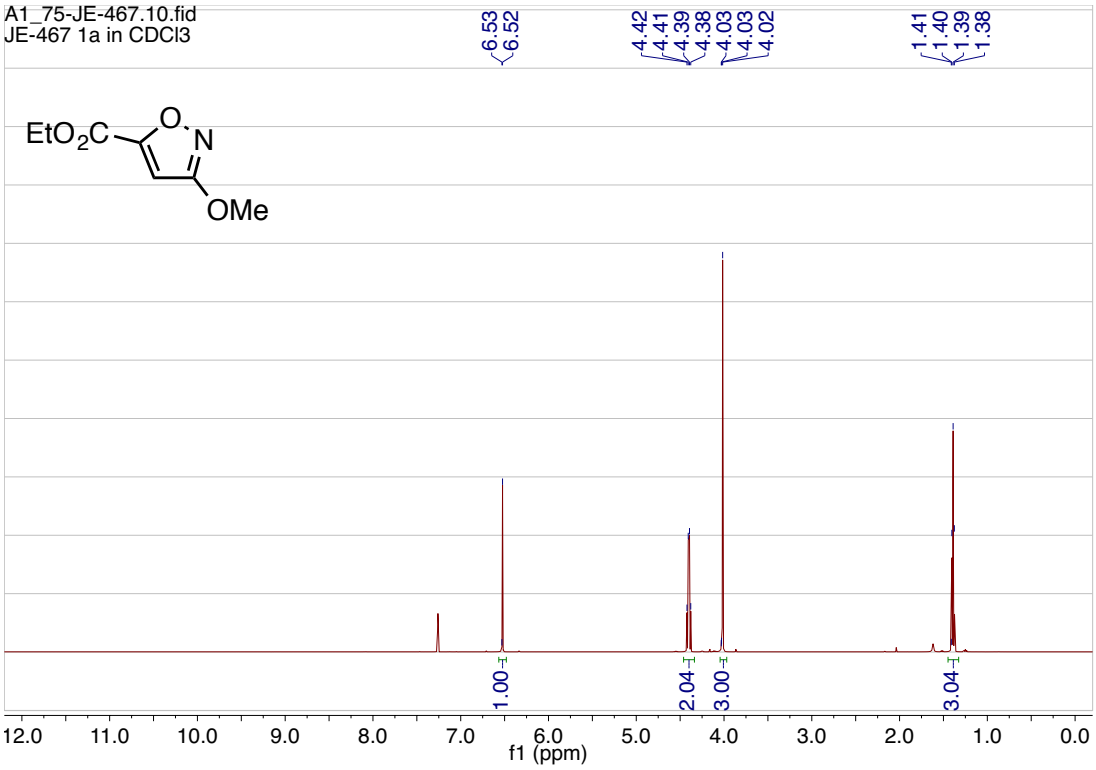
VII. NMR Spectra

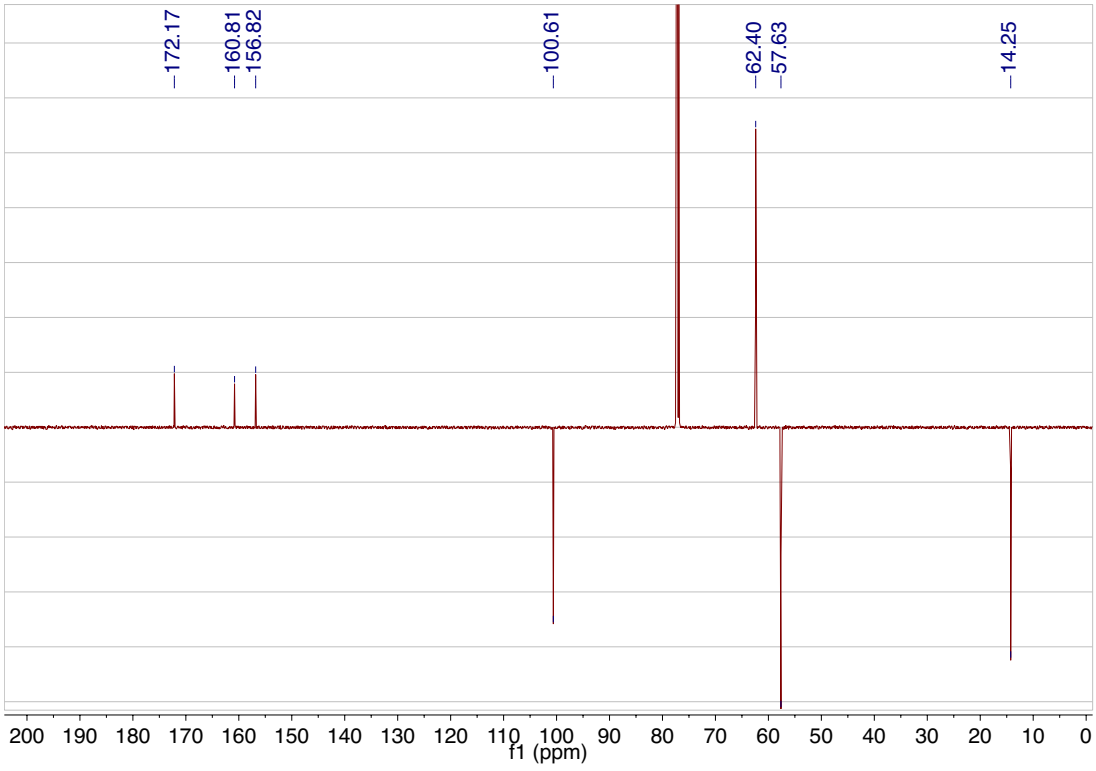


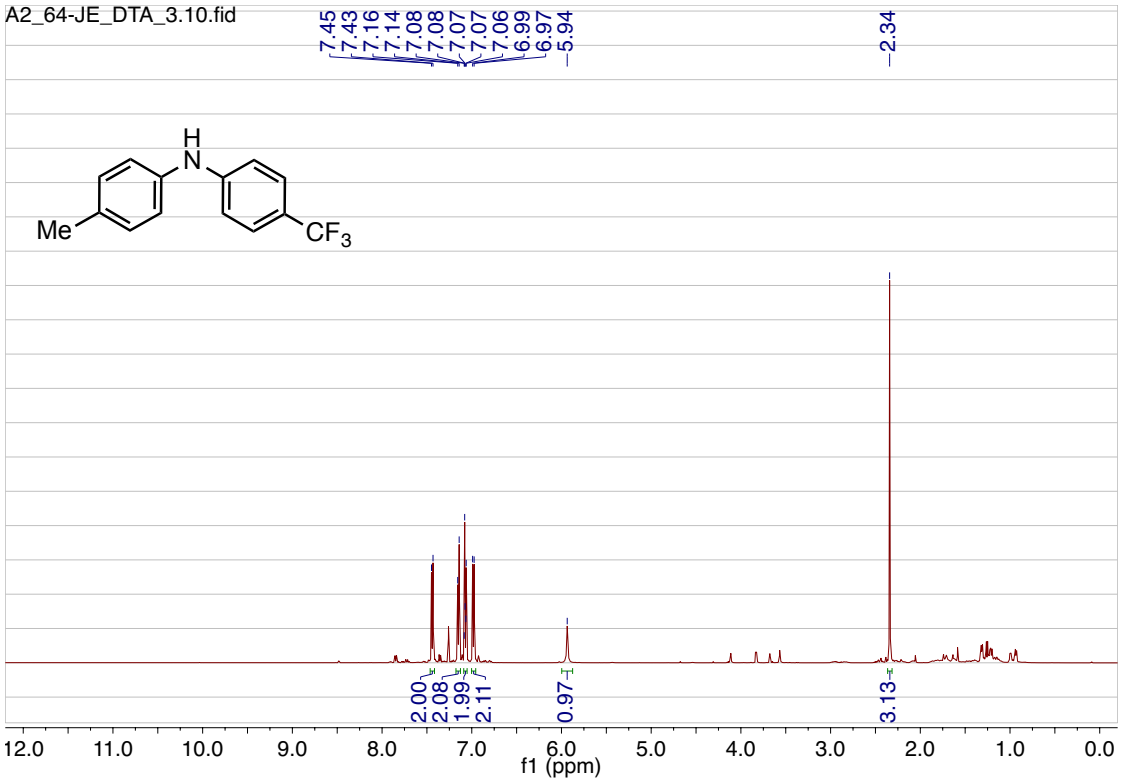


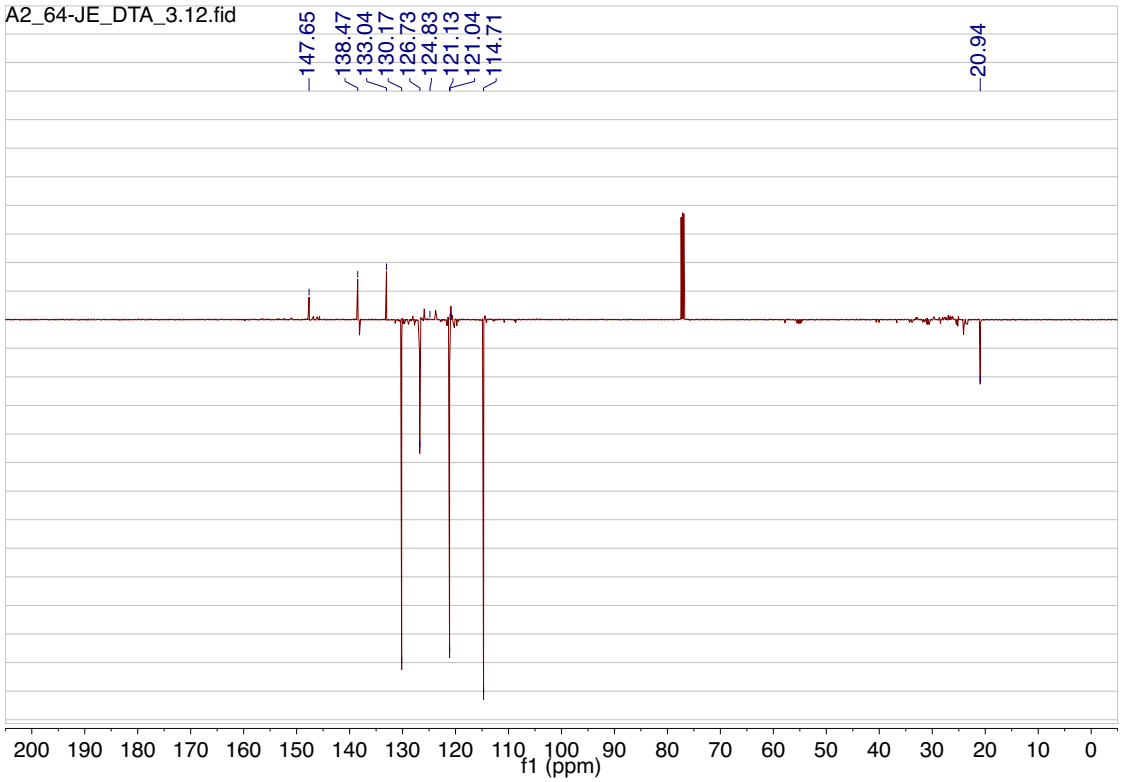

DTA-3-137_F11-21_dried.10.fid
8.09
8.08
7.38
7.22
5.91
5.91
4.50
O
N
NBn2
1.00
10.14
0.98
4.02
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
f1 (ppm)

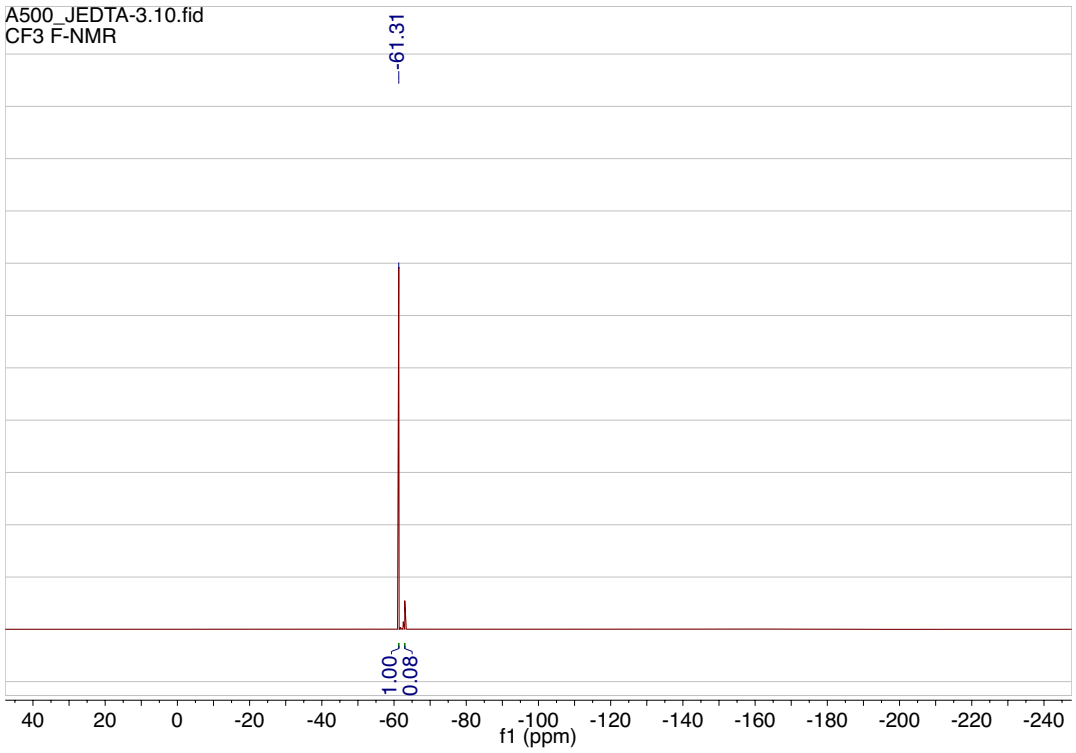

166.40
158.68
137.57
128.75
127.74
127.50
95.43
52.70
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
f1 (ppm)

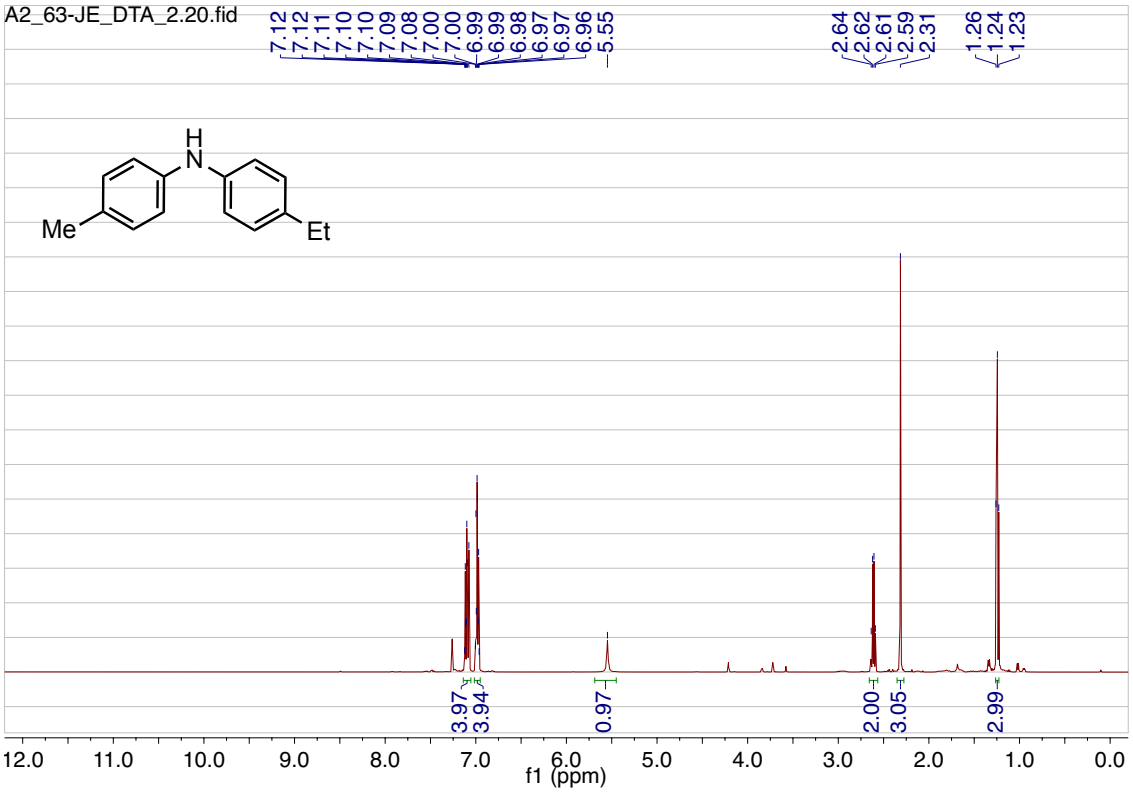

A1_75-JE-467.10.fid
JE-467 1a in CDCl3
6.53
6.52
4.42
4.41
4.39
4.38
4.03
4.03
4.02
1.41
1.40
1.39
1.38
EtO2C
O
N
OMe
1.00
2.04
3.00
3.04
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
f1 (ppm)

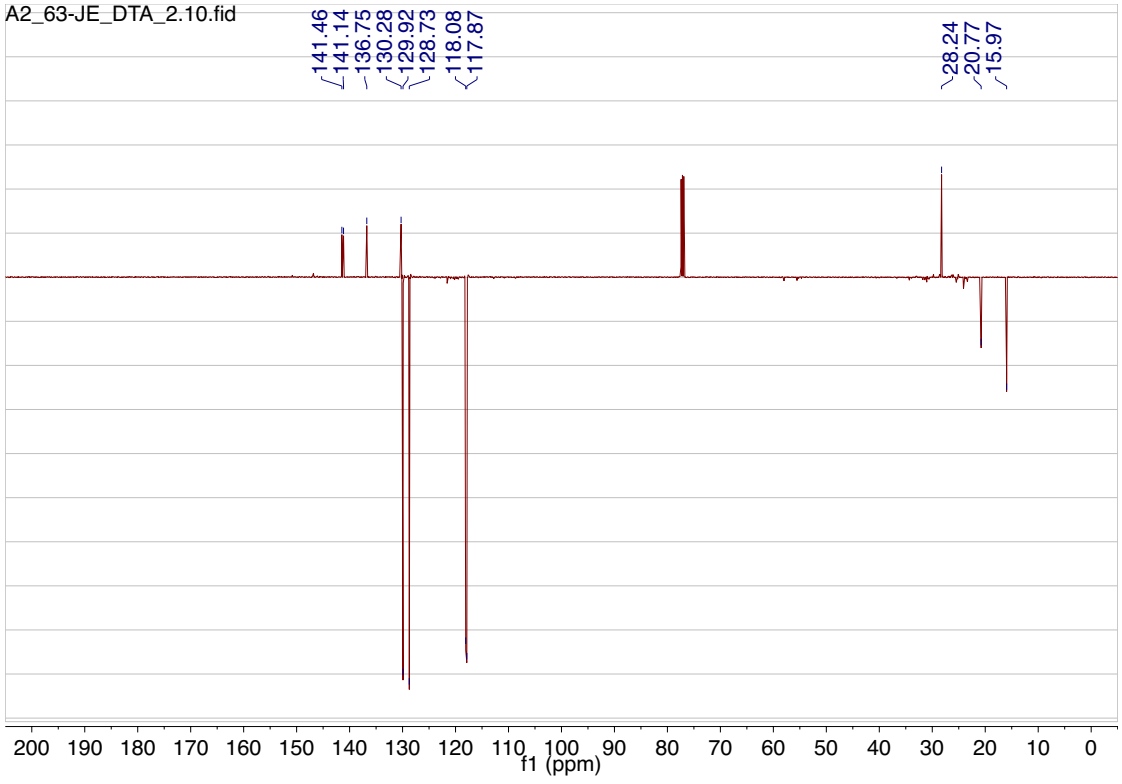

172.17
160.81
156.82
100.61
62.40
57.63
14.25
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
f1 (ppm)

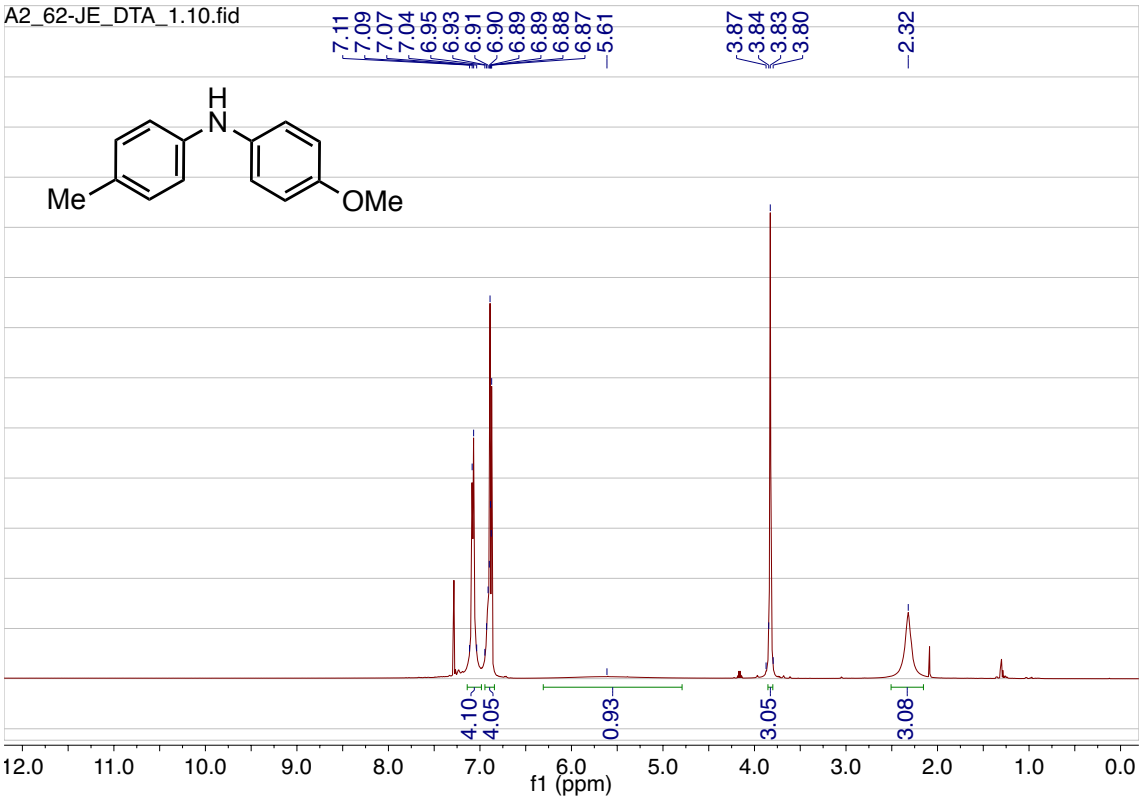

A2_64-JE_DTA_3.10.fid
7.45
7.43
7.16
7.14
7.08
7.08
7.07
7.07
7.06
6.99
6.97
5.94
2.34
H
N
Me
CF3
2.00
2.08
1.99
2.11
0.97
3.13
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
f1 (ppm)

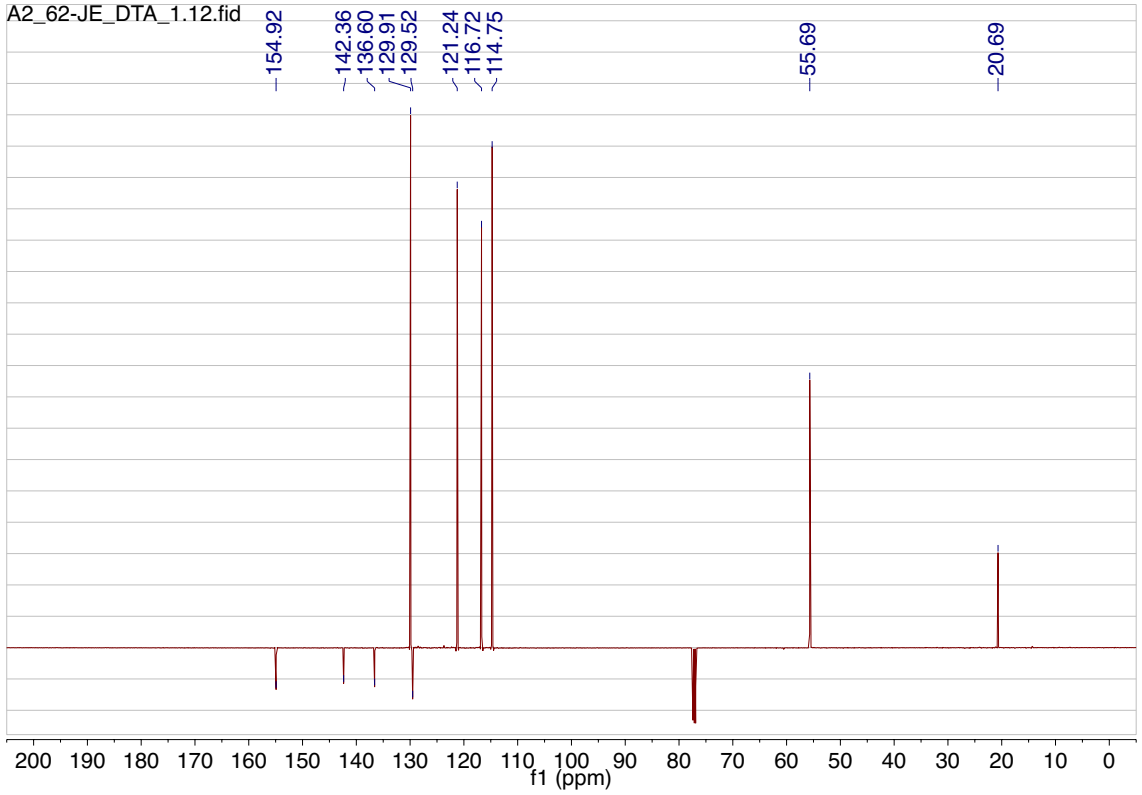

A2_64-JE_DTA_3.12.fid
147.65
138.47
133.04
130.17
126.73
124.83
121.13
121.04
114.71
20.94
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
f1 (ppm)

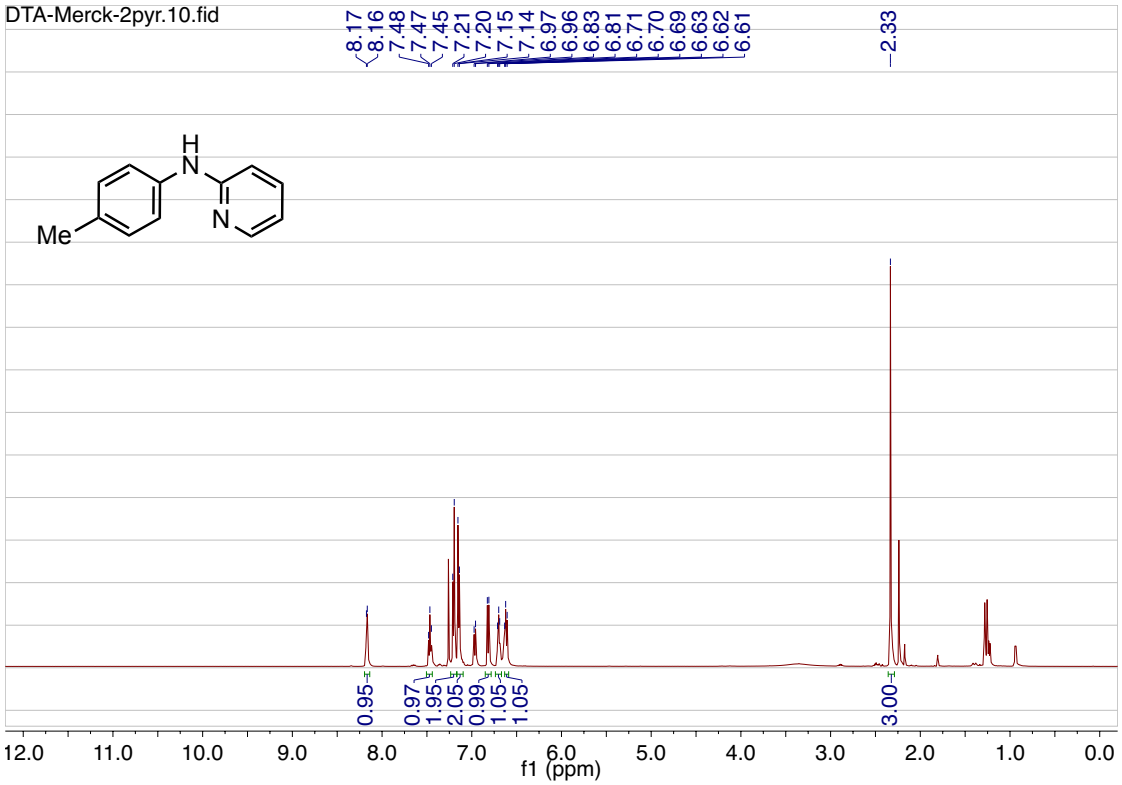

A500_JEDTA-3.10.fid
CF3 F-NMR
-61.31
1.00
0.08
40
20
0
-20
-40
-60
-80
-100
-120
-140
-160
-180
-200
-220
-240
f1 (ppm)

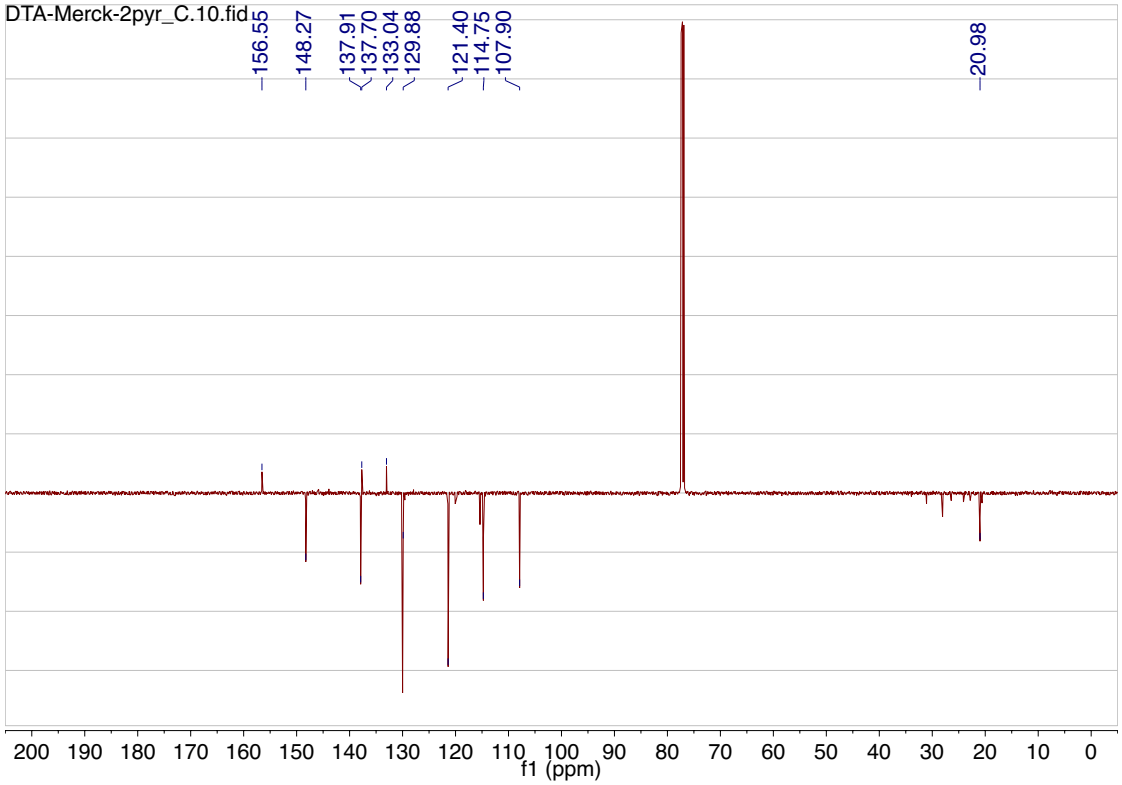

A2_63-JE_DTA_2.20.fid
7.12
7.12
7.11
7.10
7.10
7.09
7.08
7.00
7.00
6.99
6.99
6.98
6.97
6.97
6.96
5.55
2.64
2.62
2.61
2.59
2.31
1.26
1.24
1.23
H
N
Me
Et
3.97
3.94
0.97
2.00
3.05
2.99
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
f1 (ppm)

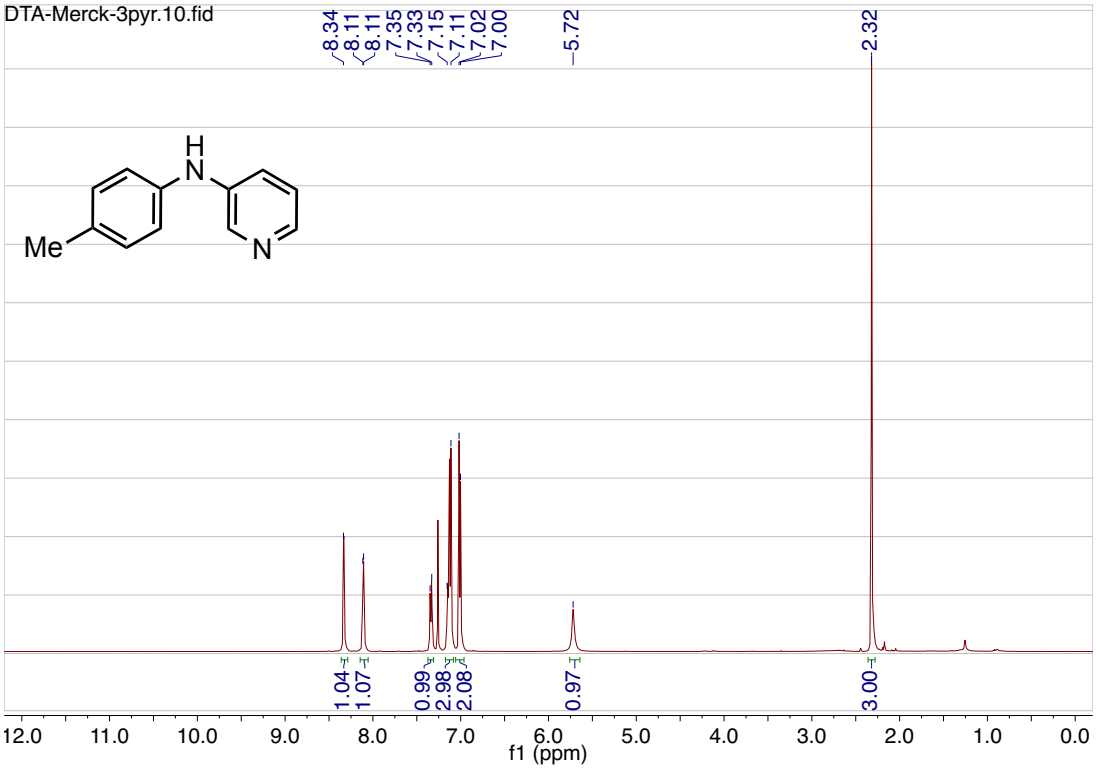

A2_63-JE_DTA_2.10.fid
141.46
141.14
136.75
130.28
129.92
128.73
118.08
117.87
28.24
20.77
15.97
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
f1 (ppm)

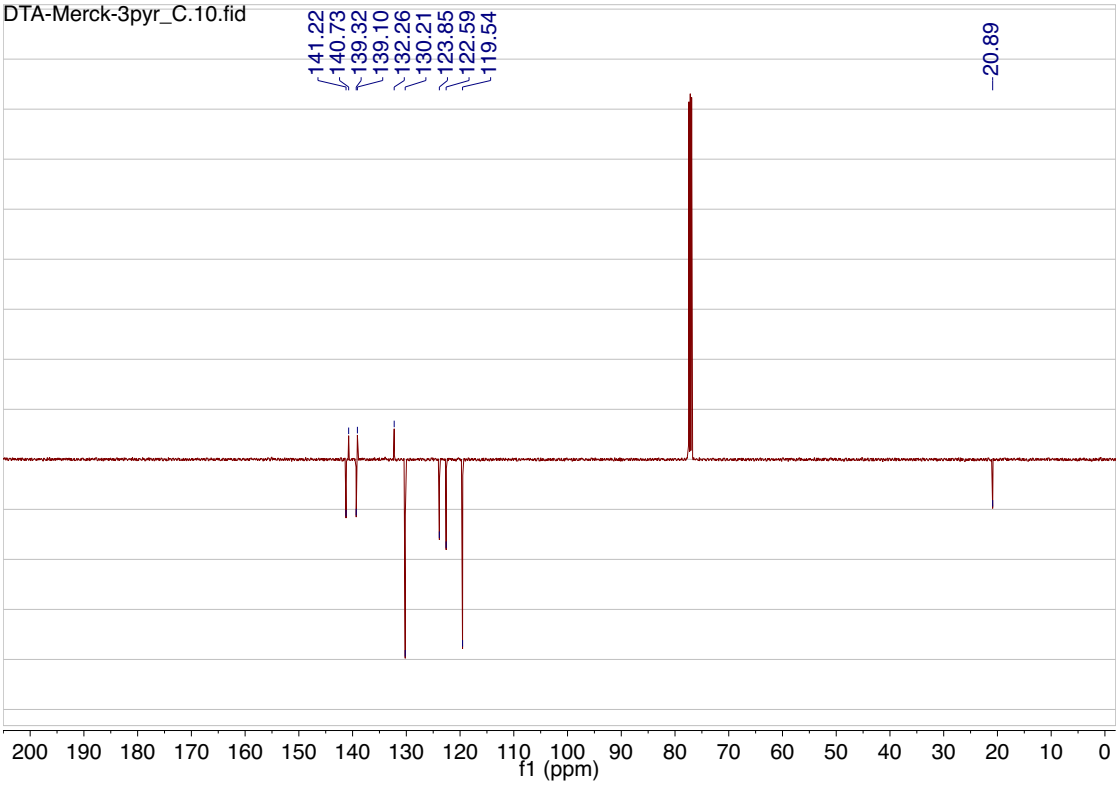

A2_62-JE_DTA_1.10.fid
7.11
7.09
7.07
7.04
6.95
6.93
6.91
6.90
6.89
6.89
6.88
6.87
5.61
3.87
3.84
3.83
3.80
2.32
H
N
Me
OMe
4.10
4.05
0.93
3.05
3.08
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
f1 (ppm)


A2_62-JE_DTA_1.12.fid
154.92
142.36
136.60
129.91
129.52
121.24
116.72
114.75
55.69
20.69
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
f1 (ppm)


DTA-Merck-2pyr.10.fid
8.17
8.16
7.48
7.47
7.45
7.21
7.20
7.15
7.14
6.97
6.96
6.83
6.81
6.71
6.70
6.69
6.63
6.62
6.61
2.33
H
N
N
Me
0.95
0.97
1.95
2.05
0.99
1.05
1.05
3.00
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
f1 (ppm)


DTA-Merck-2pyr_C.10.fid
156.55
148.27
137.91
137.70
133.04
129.88
121.40
114.75
107.90
20.98
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
f1 (ppm)


DTA-Merck-3pyr.10.fid
8.34
8.11
8.11
7.35
7.33
7.15
7.11
7.02
7.00
5.72
2.32
H
N
Me
N
1.04
1.07
0.99
2.98
2.08
0.97
3.00
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
f1 (ppm)


DTA-Merck-3pyr_C.10.fid
141.22
140.73
139.32
139.10
132.26
130.21
123.85
122.59
119.54
20.89
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
f1 (ppm)

References and Notes

1. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009.

2. M. I. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015). doi:10.1126/science.aaa8415 Medline

3. A. Lavecchia, Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov. Today* **20**, 318–331 (2015). doi:10.1016/j.drudis.2014.10.012 Medline

4. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003). doi:10.1021/ci034160g Medline

5. J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **55**, 263–274 (2015). doi:10.1021/ci500747n Medline

6. M. H. Todd, Computer-aided organic synthesis. *Chem. Soc. Rev.* **34**, 247–266 (2005). doi:10.1039/b104620a Medline

7. S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016). doi:10.1002/anie.201506101 Medline

8. M. A. Kayala, C.-A. Azencott, J. H. Chen, P. Baldi, Learning to predict chemical reactions. *J. Chem. Inf. Model.* **51**, 2209–2222 (2011). doi:10.1021/ci200207y Medline

9. J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2**, 725–732 (2016). doi:10.1021/acscentsci.6b00219 Medline

10. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **3**, 434–443 (2017). doi:10.1021/acscentsci.7b00064 Medline

11. B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* 10.1021/acscentsci.7b00303 (2017).

12. S. Kite, T. Hattori, Y. Murakami, Estimation of catalytic performance by neural network—product distribution in oxidative dehydrogenation of ethylbenzene. *Appl. Catal. A* **114**, L173–L178 (1994). doi:10.1016/0926-860X(94)80169-X

13. K. Omata, Screening of New Additives of Active-Carbon-Supported Heteropoly Acid Catalyst for Friedel-Crafts Reaction by Gaussian Process Regression. *Ind. Eng. Chem. Res.* **50**, 10948–10954 (2011). doi:10.1021/ie102477y

14. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016). doi:10.1038/nature17439 Medline

15. G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski, A. Gambin, Predicting the outcomes of organic reactions *via* machine

learning: Are current descriptors sufficient? *Sci. Rep.* **7**, 3582 (2017). doi:10.1038/s41598-017-02303-0 Medline

16. A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak, S. D. Dreher, Organic chemistry. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015). doi:10.1126/science.1259203 Medline

17. K. D. Collins, T. Gensch, F. Glorius, Contemporary screening approaches to reaction discovery and development. *Nat. Chem.* **6**, 859–871 (2014). doi:10.1038/nchem.2062 Medline

18. N. R. Draper, H. Smith, Applied Regression Analysis (Wiley, 1998).

19. M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **49**, 1292–1301 (2016). doi:10.1021/acs.accounts.6b00194 Medline

20. S. E. Denmark, N. D. Gould, L. M. Wolf, A systematic investigation of quaternary ammonium ions as asymmetric phase-transfer catalysts. Application of quantitative structure activity/selectivity relationships. *J. Org. Chem.* **76**, 4337–4357 (2011). doi:10.1021/jo2005457 Medline

21. L. P. Hammett, The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **59**, 96–103 (1937). doi:10.1021/ja01280a022

22. E. N. Bess, A. J. Bischoff, M. S. Sigman, Designer substrate library for quantitative, predictive modeling of reaction performance. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14698–14703 (2014). doi:10.1073/pnas.1409522111 Medline

23. A. Milo, E. N. Bess, M. S. Sigman, Interrogating selectivity in catalysis using molecular vibrations. *Nature* **507**, 210–214 (2014). doi:10.1038/nature13019 Medline

24. A. Milo, A. J. Neel, F. D. Toste, M. S. Sigman, Organic chemistry. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **347**, 737–743 (2015). doi:10.1126/science.1261043 Medline

25. P. Ruiz-Castillo, S. L. Buchwald, Applications of Palladium-Catalyzed C-N Cross-Coupling Reactions. *Chem. Rev.* **116**, 12564–12649 (2016). doi:10.1021/acs.chemrev.6b00512 Medline

26. P. S. Kutchukian, J. F. Dropinski, K. D. Dykstra, B. Li, D. A. DiRocco, E. C. Streckfuss, L.-C. Campeau, T. Cernak, P. Vachal, I. W. Davies, S. W. Krska, S. D. Dreher, Chemistry informer libraries: A chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chem. Sci.* **7**, 2604–2613 (2016). doi:10.1039/C5SC04751J Medline

27. E. Vitaku, D. T. Smith, J. T. Njardarson, Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among U.S. FDA approved pharmaceuticals. *J. Med. Chem.* **57**, 10257–10274 (2014). doi:10.1021/jm501100b Medline

28. K. D. Collins, F. Glorius, Intermolecular reaction screening as a tool for reaction evaluation. *Acc. Chem. Res.* **48**, 619–627 (2015). doi:10.1021/ar500434f Medline

29. M. Shahlaei, Descriptor selection methods in quantitative structure-activity relationship studies: A review study. *Chem. Rev.* **113**, 8093–8103 (2013). doi:10.1021/cr3004339 Medline

30. L. Breiman, Random Forests. *Mach. Learn.* **45**, 5–32 (2001). doi:10.1023/A:1010933404324

31. M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro, F. Borges, Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **19**, 1069–1080 (2014). doi:10.1016/j.drudis.2014.02.003 Medline

32. Another possible source of incompatibility is Pd-catalyzed isoxazole C–H arylation. However, these reactions favor electron-rich isoxazoles and typically require substantially more forcing conditions (>100°C) than the amination. See (*33*, *34*)

33. Y. Fall, C. Reynaud, H. Doucet, M. Santelli, Ligand-Free-Palladium-Catalyzed Direct 4-Arylation of Isoxazoles Using Aryl Bromides. *Eur. J. Org. Chem.* **2009**, 4041–4050 (2009). doi:10.1002/ejoc.200900309

34. M. Shigenobu, K. Takenaka, H. Sasai, Palladium-Catalyzed Direct C-H Arylation of Isoxazoles at the 5-Position. *Angew. Chem. Int. Ed.* **54**, 9572–9576 (2015). doi:10.1002/anie.201504552 Medline

35. Y. Tan, J. F. Hartwig, Palladium-catalyzed amination of aromatic C-H bonds with oxime esters. *J. Am. Chem. Soc.* **132**, 3676–3677 (2010). doi:10.1021/ja100676r Medline

36. S. Yu, G. Tang, Y. Li, X. Zhou, Y. Lan, X. Li, Anthranil: An Aminating Reagent Leading to Bifunctionality for Both C(sp^3))-H and C(sp^2))-H under Rhodium(III) Catalysis. *Angew. Chem. Int. Ed.* **55**, 8696–8700 (2016). doi:10.1002/anie.201602224 Medline

37. P. J. Bickel, B. Li, Regularization in Statistics. *Sociedad de Estadística e Investigación Operativa Test* **15**, 271–344 (2006).

38. R. Tibshirani, Regression Shrinkage and Selection *via* the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).

39. H. Zou, T. Hastie, Regularization and variable selection *via* the elastic net. *J. R. Stat. Soc. B* **67**, 301–320 (2005). doi:10.1111/j.1467-9868.2005.00503.x

40. X. Ling, Y. Xiong, R. Huang, X. Zhang, S. Zhang, C. Chen, Synthesis of benzidine derivatives via $FeCl_3 \cdot 6H_2O$-promoted oxidative coupling of anilines. *J. Org. Chem.* **78**, 5218–5226 (2013). doi:10.1021/jo4002504 Medline

41. B. Bang-Andersen, H. Ahmadian, S. M. Lenz, T. B. Stensbøl, U. Madsen, K. P. Bøgesø, P. Krogsgaard-Larsen, Structural determinants of AMPA agonist activity in analogues of 2-amino-3-(3-carboxy-5-methyl-4-isoxazolyl)propionic acid: Synthesis and pharmacology. *J. Med. Chem.* **43**, 4910–4918 (2000). doi:10.1021/jm0003586 Medline

42. S. S. Kampmann, A. N. Sobolev, G. A. Koutsantonis, S. G. Stewart, Stable Nickel (0) Phosphites as Catalysts for C—N Cross-Coupling Reactions. *Adv. Synth. Catal.* **356**, 1967–1973 (2014). doi:10.1002/adsc.201400201

43. Y. Zhao, B. Huang, C. Yang, B. Li, B. Gou, W. Xia, Photocatalytic Cross-Dehydrogenative Amination Reactions between Phenols and Diarylamines. *ACS Catal.* **7**, 2446–2451 (2017). doi:10.1021/acscatal.7b00192