Electronic Supplementary Material (ESI) for Reaction Chemistry & Engineering.
This journal is © The Royal Society of Chemistry 2022

# The Effect of Chemical Representation on Active Machine Learning Towards Closed-Loop Optimization

A. Pomberger[1], A. A. Pedrina McCarthy[2], A. Khan[1], S. Sung[3], C. J. Taylor[1,4], M. J. Gaunt[2], L. Colwell[2], D. Walz[5] and A. Lapkin[1,3]

[1]Department of Chemical Engineering and Biotechnology, Cambridge, CB3 0AS, United Kingdom
[2]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom
[3]Cambridge Centre for Advanced Research and Education in Singapore Ltd., CREATE Tower 05-05, 138602 Singapore
[4]Astex Pharmaceuticals, 436 Cambridge Science Park, Milton, Cambridge CB4 0QA, United Kingdom
[5]BASF SE Data Science for Materials, Carl-Bosch-Strasse 38, 67056 Ludwigshafen am Rhein, Germany

## Supporting information

## Table of Contents

# General Considerations

Unless mentioned otherwise, all solvents, reagents, and substrates were purchased from commercial suppliers and were used as received, including 2’-deoxyguanosine (Fluorochem), 3-nitropyridine (Sigma-Aldrich), $Ru(bpy)_3(PF_6)_2$ (Sigma-Aldrich), DMSO (Fisher), piperidine (Sigma-Aldrich), and all photocatalysts, additives and amines tested in the high-throughput experiments (various suppliers). Compound names are those generated by ChemDraw 16.0 software (PerkinElmer), following the IUPAC nomenclature.

## Analytical Methods

**Proton nuclear magnetic resonance** ($^1$H NMR) spectra were recorded at ambient temperature on a Bruker Avance III HD spectrometer (400 MHz), a Bruker Avance III HD Smart Probe spectrometer (500 MHz) or a Bruker Avance II+ spectrometer (700 MHz). Chemical shifts (δ) were reported in ppm and quoted to the nearest 0.01 ppm relative to the residual protons in $CDCl_3$ (7.26 ppm), $D_2O$ (4.79 ppm) and DMSO-$d_6$ (2.05 ppm) with coupling constants (J) were quoted in Hertz (Hz). Coupling constants were quoted to the nearest 0.1 Hz and multiplicity reported according to the following convention: s = singlet, d = doublet, t = triplet, q = quartet, qnt = quintet, sxt = sextet, spt = septet, oct = octet, m = multiplet, br = broad and associated combinations, e.g. dd = doublet of doublets. Where coincident coupling constants have been observed, the apparent (app) multiplicity of the proton resonance has been reported. Data were reported as follows: chemical shift (multiplicity, coupling constants, number of protons and molecular assignment).

**Carbon nuclear magnetic resonance** ($^{13}$C NMR) spectra were recorded at ambient temperature on a 400 MHz Bruker Avance III HD spectrometer (101 MHz) or a 500 MHz Bruker Avance III HD Smart Probe spectrometer (126 MHz). Chemical shifts (δ) were reported in ppm and quoted to the nearest 0.1 ppm relative to the residual solvent peaks in $CDCl_3$ (77.16 ppm) and DMSO-$d_6$ (39.52 ppm). DEPT-135, NOE experiments and 2D experiments (COSY, HMBC and HSQC) were used to support assignments when appropriate but were not included herein. Fluorine nuclear magnetic resonance ($^{19}$F NMR) spectra were recorded at ambient temperature on a 400 MHz Bruker Avance III HD spectrometer (376 MHz). Chemical shifts (δ) were reported in ppm and quoted to the nearest 0.1 ppm relative to the residual solvent peaks in $CDCl_3$ (77.16 ppm).

**Infrared (IR) spectra** were collected using a Thermo Fisher Scientific Nicolet Summit Pro equipped with an Everest ATR, with absorption maxima ($\nu_{max}$) quoted in wavenumbers ($cm^{-1}$).

**Analytical thin layer chromatography** (TLC) was performed using pre-coated Merck glass-backed silica gel plates (Silica gel 60 F254 0.2 mm). Visualization was achieved using ultraviolet light (254 nm) and chemical staining with basic potassium permanganate solution as appropriate, or otherwise stated. Flash column chromatography was undertaken on Fluka or Material Harvest silica gel (230-400 mesh) under a positive pressure of air unless otherwise stated.

**Analytical mobile phases for LC–MS** in both projects were A = 2.5 L acetonitrile + 131 mL water + 1.25 mL and formic acid, B = 2.4 L water + 1.50 g ammonium formate + 2.4 mL formic acid. The autosampler was washed between each run with a 1:1 mixture of acetonitrile:water. Gradients were generally 5-95% over 0.8/1.2 min.

**Low resolution LC–MS for HT quantification:** samples were analysed using a Shimadzu LC–MS; SIL-20AC XR autosampler, 2 × LC-20 AD XR pumps, CBM-20A communicator, SPD-M20A photodiode array (PDA), CTO-20AC column oven and LCMS-2020 mass spectroscopy unit. The 384-well analysis plate was placed into autosampler on a Shimadzu microtiter plate (MTP) rack. All samples were run on a Kinetex® 2.6 µm, 50 × 2.1 mm, 100 Å C18 column (Cat. No. H16-189446). The mass spectrometry unit was set to dual mode (DUIS), in which both atmospheric pressure chemical ionisation (APCI) and electrospray ionisation (ESI) mode are used simultaneously, in the positive mode and set for selective ion monitoring of M+1 for product and internal standard (scan speed 15000u). Data analysis was undertaken using Shimadzu Lab Solutions software (Version 5.97) and exported into Microsoft Excel for further statistical tests and data visualisation.

**Nanoscale C–H activation:** Nanoscale reactions (50-100 nmol) were run using Corning 1,536-well plates (Corning Echo qualified, Cat. No. 3730, Cyclic Olefin-Copolymer COC, 12.5 µL-wells, flat bottom, clear) as reaction plates. Reactions at elevated temperatures were ran in Corning 1,536-well White High Base plates (Cyclic Olefin Copolymer Cat. No. 4570) and typically with Axygen 384-well plates (Cat No. P-284-120SQ-C, Polypropylene, 120 µL, V-bottom, translucent) used as solution source plates for stock solutions and for analytical plates on LC–MS equipment. Analysis plates were sealed with gas permeable adhesive sheets (4titude, Cat. No. 4ti-0516/384).

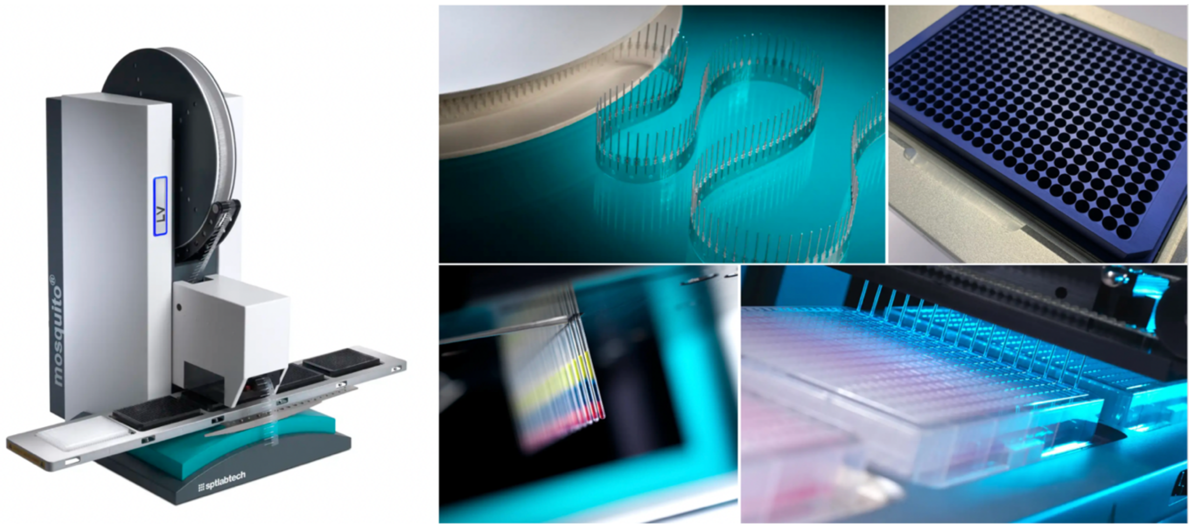
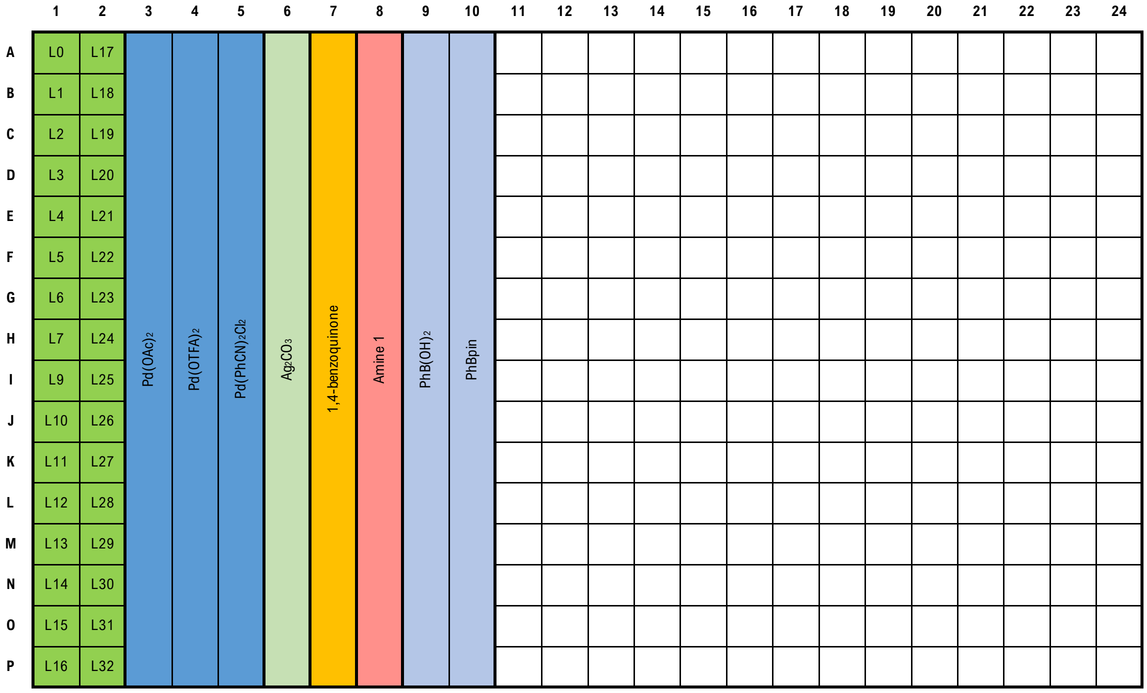


*Figure S1 The Mosquito liquid handling robot*

**For reactions at elevated temperatures:** the 1,536-well plates were covered by a perfluoroalkoxy alkane (PFA) mat (0.125 mm thickness, FLONFILM™ 600 PFA film), followed by a neoprene rubber matt (on top and below) and then secured within a custom-built plate-sealing device, which was tightened through gradual even turning of all 14 screws in a crosswise pattern. The entire assembly was heated in an oven for the reaction duration. Once complete the assembly is cooled in a laboratory fridge to ~10 °C, minimising contamination when unsealed. Following this, the plate is removed from the assembly and centrifuged prior to Mosquito dosing, Figure S1 illustrates the Mosquito liquid handling robot. However, commercial plate-sealing alternatives to this are now available from Analytical Sales and Services Inc. (Cat. No. 1626100).

## High Throughput Experimentation

**Reaction preparation:** stock solutions of the reaction components were prepared in *N*-methyl-pyrrolidone (NMP) according to table S1, below. Stock solutions were then charged into a 384-well source plate according to the source plate layouts in Figure S2. The required volume for each source well was calculated to include an additional 20 µL top-up, ensuring that there would be an excess during plate dosing.



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | L0 | L17 | $Pd(OAc)_2$ | $Pd(OTFA)_2$ | $Pd(PhCN)_2Cl_2$ | $Ag_2CO_3$ | 1,4-benzoquinone | Amine 1 | $PhB(OH)_2$ | PhBpin | | | | | | | | | | | | | | |
| B | L1 | L18 | | | | | | | | | | | | | | | | | | | | | | |
| C | L2 | L19 | | | | | | | | | | | | | | | | | | | | | | |
| D | L3 | L20 | | | | | | | | | | | | | | | | | | | | | | |
| E | L4 | L21 | | | | | | | | | | | | | | | | | | | | | | |
| F | L5 | L22 | | | | | | | | | | | | | | | | | | | | | | |
| G | L6 | L23 | | | | | | | | | | | | | | | | | | | | | | |
| H | L7 | L24 | | | | | | | | | | | | | | | | | | | | | | |
| I | L9 | L25 | | | | | | | | | | | | | | | | | | | | | | |
| J | L10 | L26 | | | | | | | | | | | | | | | | | | | | | | |
| K | L11 | L27 | | | | | | | | | | | | | | | | | | | | | | |
| L | L12 | L28 | | | | | | | | | | | | | | | | | | | | | | |
| M | L13 | L29 | | | | | | | | | | | | | | | | | | | | | | |
| N | L14 | L30 | | | | | | | | | | | | | | | | | | | | | | |
| O | L15 | L31 | | | | | | | | | | | | | | | | | | | | | | |
| P | L16 | L32 | | | | | | | | | | | | | | | | | | | | | | |

*Figure S2 Source plate layout for the high throughput optimization*

The Mosquito liquid handling robot was used to sequentially transfer 0.5 µL aliquots of each of the six reaction components from the 384-well source plate to the 1,536-well reaction plate (dosing sequence given in Table S1). Upon dosing the final reagent, three cycles of the Mosquito's dispense mix setting (500 nL, move 0.5 mm), was used to ensure all reagents were evenly distributed. Silver carbonate is completely insoluble in NMP and settles at the bottom of source plate wells in ca. 1 min, blocking pipettes and leading to inconsistent stoichiometries. As such, preparation of the source and reaction plates required an alternate method; i) a slurry of the required concentration in NMP was prepared; ii) under vigorous stirring, using hand-held electronic pipettes, the source plate was charged with the required aliquot (68 µL); iii) the Mosquito was paused after each transfer from source plate to the reaction plate and wells containing silver carbonate were mixed with the 'aspirate-mix' function (20 µL, 3 rounds) of a multi-channel electronic pipette. This method allowed for consistent dispensation of insoluble reagents.

*Table S1 Dosing table of the HT optimization of C(sp3)–H activation of amine* ***1***

| **Dosing sequence** | **Reagent** | **Equiv.** | **Concentration / м** | **Min source plate vol. (+ top up) / µL** | **Aliquot / µL** |
|---|---|---|---|---|---|
| 1 | MPAA ligands | 0.25 | 0.060 | 24 (+20) | 0.5 |
| 2 | Pd pre-catalyst | 0.10 | 0.024 | 16 (+20) | 0.5 |
| 3 | $Ag_2CO_3$ | 2.5 | 0.60 | 48 (+20) | 0.5 |
| 4 | 1,4-benzoquinone | 2.0 | 0.48 | 48 (+20) | 0.5 |
| 5 | amine | 2.5 | 0.60 | 48 (+20) | 0.5 |
| 6 | boronates | 1.0 | 0.24 | 24 (+20) | 0.5 |
| - | **Total** | - | 0.04 | - | 3.0 |

The reaction plate was then placed into a custom-designed aluminium plate sealer (commercial alternatives now available, i.e., Analytical Sales NanoNest), topped with a chemically inert PFA film (0.125 mm thick) and a silicone gasket. The assembly was secured by gradually tightening 14 screws in a cross-wise pattern, ensuring even compression on all sides of the reaction plate. The entire assembly was then placed in a temperature-controlled laboratory oven set to 55 °C.

**Analysis:** After the desired reaction time had elapsed, the assembly was removed from the oven, allowed to cool to room temperature and placed in a 10 °C fridge for 10 min prior to opening (generating negative pressure inside the wells and avoiding messy pressure release). The Mosquito was used to aspirate and transfer 100 nL from each reaction well into a 384-well 'analysis plate' which was pre-loaded with 50 µL of a quenching diluent, MeCN:$H_2O$:formic acid (2:1:1) that contained a known concentration (0.04 mM) of an internal standard (*N*,*N*-dibenzylaniline, DBA). This plate was then diluted further by the addition of 50 µL of the IS-doped diluent, sealed with an adhesive LC–MS autosampler-compatible sealing film and analysed by LC–MS (Shimadzu LC–MS-2020, selective ion monitoring to follow the total ion count of the IS and arylated product). Prior to calibrant and reaction sample analyses, 200 matrix-matched ‘sacrificial samples’ were used to pre-condition the LC–MS, ensuring consistent performance between runs.

**Calibration samples:** Using authentic, independently synthesized, product four known-concentration calibration samples were prepared with product/IS ratios of 0.25, 0.50, 0.75 and 1.0 (Table S2), matching the concentration of the reaction samples. These were analysed immediately preceding and following the reaction samples, repeats were averaged to build a calibration curve for the desired product (Figure S3). Analytical data were pre-processed in the native LabSolutions software (version 5.97), before being transferred to Microsoft Excel for final processing and visualisation.

*Table S2 Composition of calibration samples*

| Calibration Sample | IS conc. / mM | Product Conc. / mM | Calib. (% Yield) | Prod/IS TIC Ratio |
|---|---|---|---|---|
| 1 | 0.040 | 0.000 | 0 | 0.00 |
| 2 | 0.040 | 0.010 | 25 | 0.26 |
| 3 | 0.040 | 0.020 | 50 | 0.50 |
| 4 | 0.040 | 0.030 | 75 | 0.71 |
| 5 | 0.040 | 0.040 | 100 | 0.97 |
| Reaction Sample | 0.040 | Unknown | Unknown | Unknown |

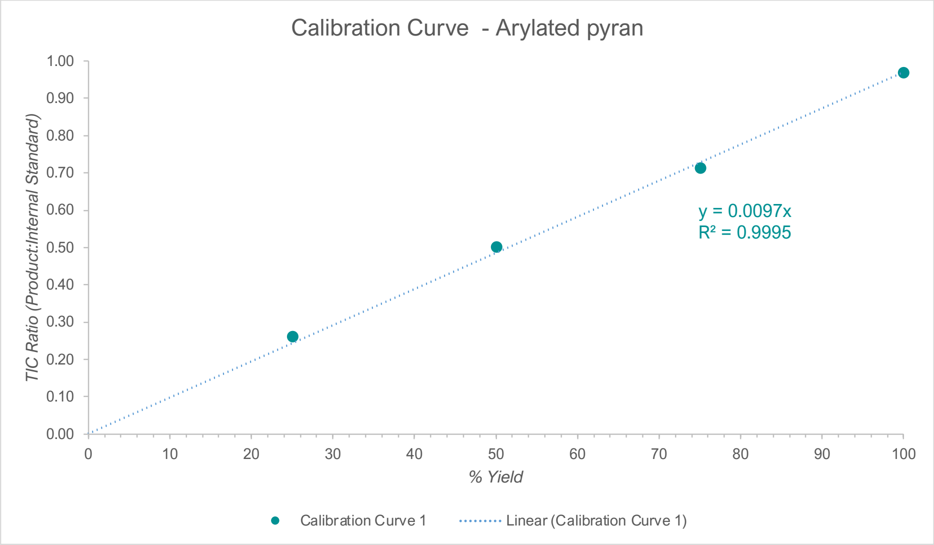
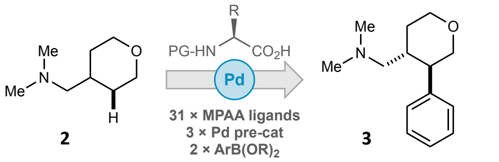
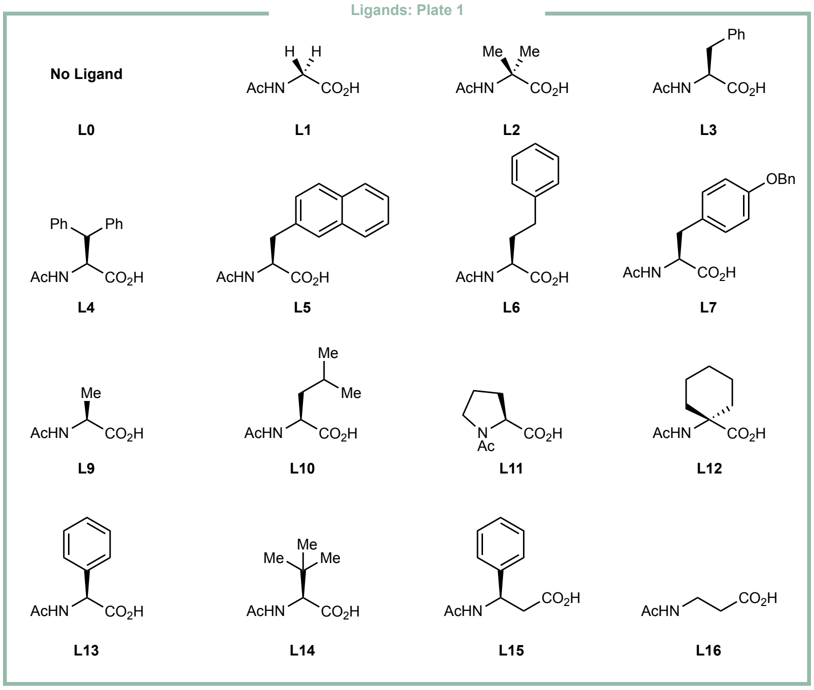


*Figure S3 Product calibration curve used to quantify product in the reaction mixture*

## Reaction Scheme and Ligand Structures





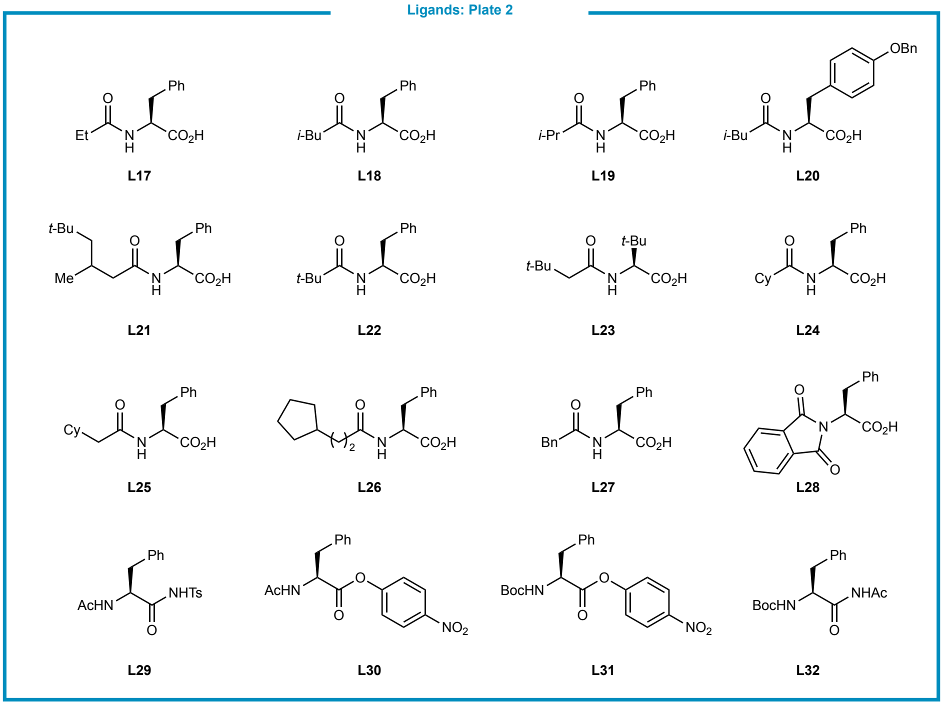
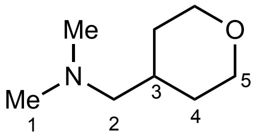


*Figure S4 Ligands explored in the HTE C(sp3)–H activation of tertiary alkylamine **2**. Ligand 8 is not represented in the table since this was initially the blank – for convenience the blank was changed as L0.*

## Synthesis of Materials

### *N,N-dimethyl-1-(tetrahydro-2H-pyran-4-yl)methanamine (**2**)*



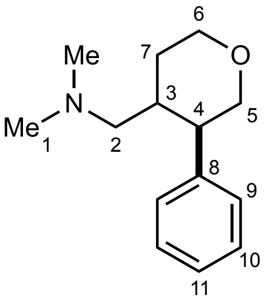
To (tetrahydro-*2H*-pyran-4-yl)methanamine (2.6 mL, 21.9 mmol) was added formaldehyde (5.6 mL, 37% aq., 69.0 mmol) at 0 °C with vigorous stirring. Formic acid (4.7 mL, 91.0 mmol) was added over 5 minutes and the reaction left to rise to room temperature over 1 hour, then heated to 85 °C for 24 hours. The yellow reaction mixture was cooled to room temperature and HCl (20 mL, 3.0 M aq.) was added. The aqueous mixture was washed with diethyl ether (3×40 mL) and the pH adjusted ca. pH 8. The aqueous layer was then extracted with diethyl ether (3×40 mL). The combined organic extracts were dried over $MgSO_4$, filtered and concentrated *in vacuo*. The resulting colourless oil was purified by vacuum distillation (93-95 °C, 75 mbar) to yield **2** as a colourless oil (1.45 g, 46%).

**$^1$H NMR** (400 MHz, $CDCl_3$) δ 3.96 (dd, *J* = 11.3, 3.7 Hz, 2H, $H_{5eq}$), 3.38 (td, *J* = 11.8, 2.0 Hz, 2H, $H_{5ax}$), 2.20 (s, 6H, $H_1$), 2.11 (d, *J* = 6.9 Hz, 2H, $H_2$), 1.80 – 1.59 (m, 3H, $H_3$/$H_{4eq}$), 1.26 (qd, *J* = 12.0, 4.5 Hz, 1H, $H_{4ax}$).

**$^{13}$C NMR** (101 MHz, $CDCl_3$) δ 67.9 ($C_5$), 66.4 ($C_2$), 45.9 ($C_1$), 33.2 ($C_3$), 31.6 ($C_4$).

Analytical data agrees with those reported previously.[1]

### *N,N-dimethyl-1-(3-phenyltetrahydro-2H-pyran-4-yl)methanamine (**3**)*



An oven dried 10 mL microwave vial, equipped with a stir bar was charged with $Pd(OAc)_2$ (6.7 mg, 0.03 mmol), *N*-acyl-*L*-*tert*-leucine (10.4 mg, 0.06 mmol), benzoquinone (64.9 mg, 0.60 mmol), silver carbonate (207 mg, 0.75 mmol), *N,N*-dimethyl-1-(tetrahydro-2*H*-pyran-4-yl)methanamine (107 mg, 0.75 mmol) and NMP (6.5 mL). The vial was sealed, heated to 50 °C and stirred at 1000 rpm before benzeneboronic acid (36.6 mg, 0.30 mmol) was added as a solution in NMP (1 mL). The reaction mixture was stirred for 18 h, cooled to room temperature, diluted with diethyl ether (50 mL) and washed with 1% aq. NaOH (5×200 mL). The organic phase was dried over $MgSO_4$, filtered, and concentrated *in vacuo* to yield a brown oil which was purified by column chromatography (silica gel, 99:1 $CH_2Cl_2$:0.45 M $NH_3$ in MeOH, $R_f$ = 0.25) to yield **3** as a light brown oil (22.5 mg, 34%).

**$^1$H NMR** (400 MHz, $CDCl_3$) δ 7.31 (t, *J* = 7.3 Hz, 2H, $H_{10}$), 7.26 – 7.20 (m, 1H, $H_{11}$), 7.17 (d, *J* = 6.8 Hz, 2H, $H_9$), 4.09 (dd, *J* = 11.5, 4.2 Hz, 1H, $H_{6eq}$), 3.87 (dd, *J* = 11.4, 4.4 Hz, 1H, $H_{5eq}$), 3.53 (td, *J* = 11.9, 2.1 Hz, 1H, $H_{6ax}$), 3.35 (t, *J* = 11.2 Hz, 1H, $H_{5ax}$), 2.51 (td, *J* = 10.8, 4.4 Hz, 1H, $H_4$), 2.10 (s, 6H, $H_1$), 2.08 – 1.95 (m, 3H, , $H_{7eq}$ , $H_3$, $H_{2'}$), 1.89 (dd, *J* = 11.6, 2.5 Hz, 1H, $H_{2''}$), 1.43 (tdd, *J* = 13.4, 10.6, 4.5 Hz, 1H, $H_{7ax}$);

**$^{13}$C NMR** (101 MHz, $CDCl_3$) δ 140.6 ($C_8$), 128.6 ($C_{10}$), 127.9 ($C_9$), 126.8 ($C_{11}$), 73.8 ($C_5$), 68.4 ($C_6$), 63.6 ($C_2$), 48.3 ($C_4$), 45.8 ($C_1$), 38.0 ($C_3$), 31.2 ($C_7$);

**IR** $ν_{max}$/cm$^{-1}$ (thin film) 2945, 2817, 2763, 1602, 1493, 1455, 1455, 1385, 1301, 1266, 1236, 1177, 1129, 1088, 1052, 1041, 1013, 997, 977, 901, 869, 846, 793;

**HRMS** (m/z): [M]+ calcd for $C_{14}H_{21}NO$, 220.1696; found, 220.1698.

## Preparation of the Dataset

The obtained HTE data (see heatmap in Figure S5) was analyzed with respect to the deviation of the single measurements (within the quartet) and eventually used to create the dataset for modelling. Within the heatmap the lines refer to the number of the ligand (L0 = no ligand) and the columns refer to the precatalyst (C1 = palladium(II)acetate, C2 = palladium(II)trifluoroacetate, C3 = bis(benzonitrile)palladium(II)chloride) as well as the boronates. As visible, all experiments were conducted four times and the mean yield of each quartet was used for the modelling. We obtained an average standard error of the mean of 2.8% (Eqn. 3), an average mean absolute deviation of 3.1% (Eqn. 4) and an average maximal deviation of 5.3% (Eqn. 5).

$$\sigma_{x_j} = \sqrt{\frac{1}{n-1}\sum_{i}^{n}\left(x_{ij} - \bar{x}_j\right)^2}$$

Eqn. 1 (standard deviation)

$$\sigma_{\bar{x}_j} = \frac{1}{\sqrt{n}}\sigma_{x_j}$$

Eqn. 2 (standard error of the mean)

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{N}\sum_j^N \sigma_{\bar{x}_j}^2}$$

Eqn. 3 (average standard error of mean)

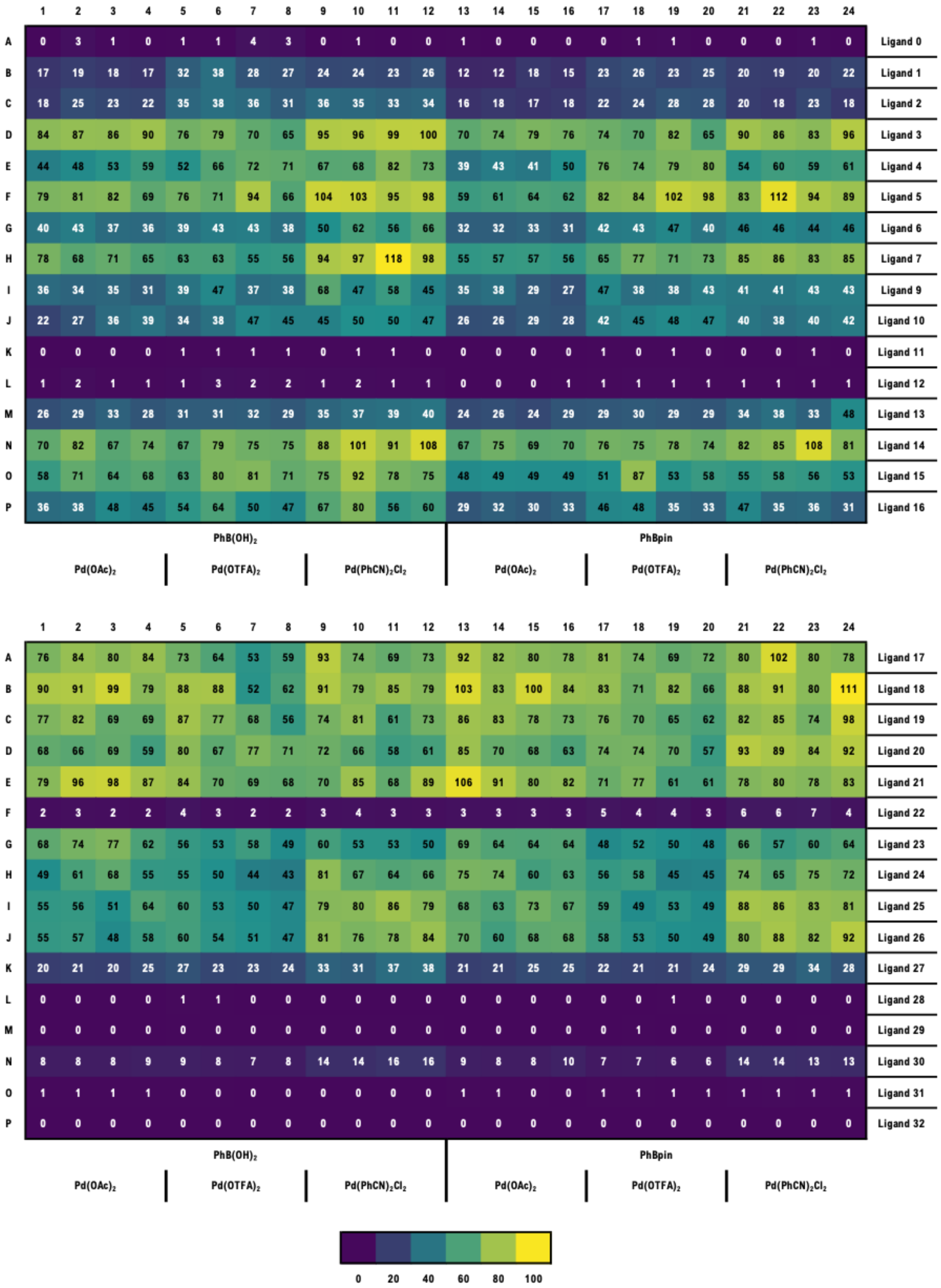
$$MAD = \frac{1}{N}\sum_j^N \frac{1}{n-1}\sum_i^n \left|x_{ij} - \bar{x}_j\right|$$

Eqn. 4 (average mean absolute deviation)

$$maxAD = \frac{1}{N}\sum_j^N \max_i \left|x_{ij} - \bar{x}_j\right|$$

Eqn. 5 (average max absolute deviation)

Where $n = 4$ is the number of repetitions and $N = 186$ is the number of conditions.



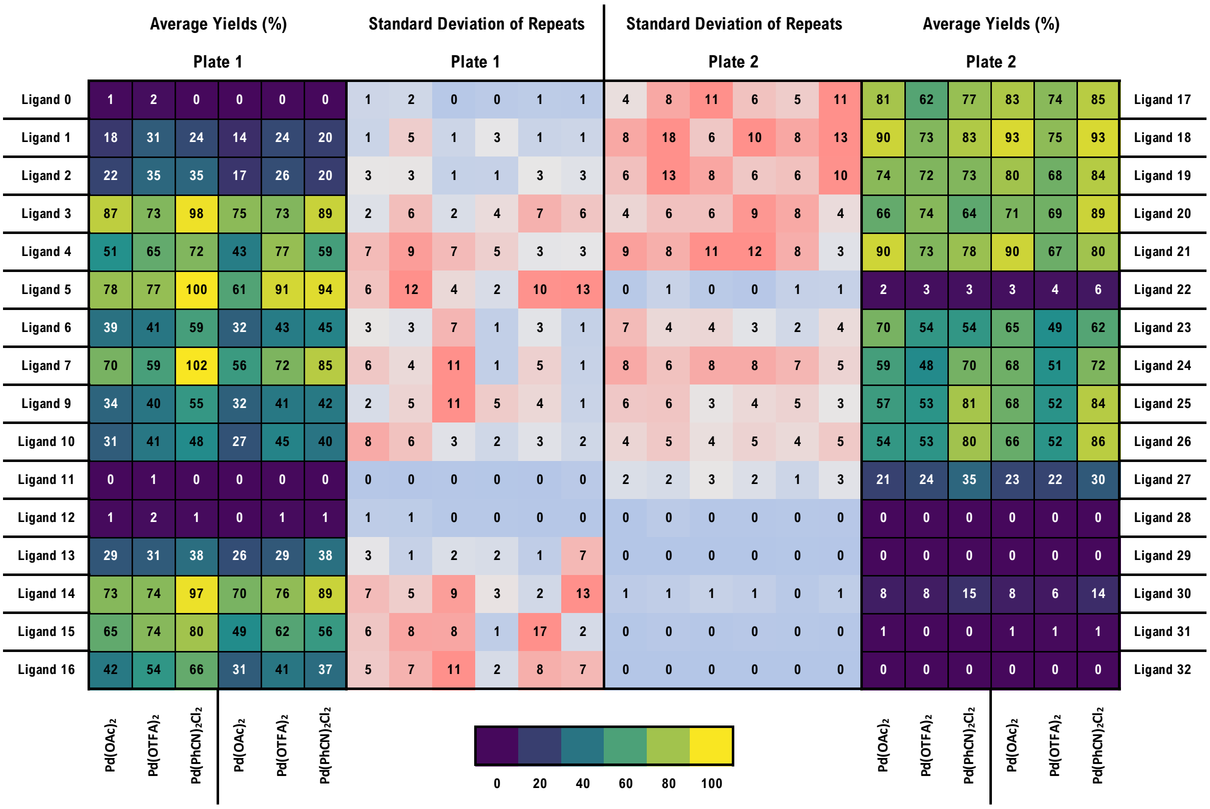
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 3 | 1 | 0 | 1 | 1 | 4 | 3 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | Ligand 0 |
| B | 17 | 19 | 18 | 17 | 32 | 38 | 28 | 27 | 24 | 24 | 23 | 26 | 12 | 12 | 18 | 15 | 23 | 26 | 23 | 25 | 20 | 19 | 20 | 22 | Ligand 1 |
| C | 18 | 25 | 23 | 22 | 35 | 38 | 36 | 31 | 36 | 35 | 33 | 34 | 16 | 18 | 17 | 18 | 22 | 24 | 28 | 28 | 20 | 18 | 23 | 18 | Ligand 2 |
| D | 84 | 87 | 86 | 90 | 76 | 79 | 70 | 65 | 95 | 96 | 99 | 100 | 70 | 74 | 79 | 76 | 74 | 70 | 82 | 65 | 90 | 86 | 83 | 96 | Ligand 3 |
| E | 44 | 48 | 53 | 59 | 52 | 66 | 72 | 71 | 67 | 68 | 82 | 73 | 39 | 43 | 41 | 50 | 76 | 74 | 79 | 80 | 54 | 60 | 59 | 61 | Ligand 4 |
| F | 79 | 81 | 82 | 69 | 76 | 71 | 94 | 66 | 104 | 103 | 95 | 98 | 59 | 61 | 64 | 62 | 82 | 84 | 102 | 98 | 83 | 112 | 94 | 89 | Ligand 5 |
| G | 40 | 43 | 37 | 36 | 39 | 43 | 43 | 38 | 50 | 62 | 56 | 66 | 32 | 32 | 33 | 31 | 42 | 43 | 47 | 40 | 46 | 46 | 44 | 46 | Ligand 6 |
| H | 78 | 68 | 71 | 65 | 63 | 63 | 55 | 56 | 94 | 97 | 118 | 98 | 55 | 57 | 57 | 56 | 65 | 77 | 71 | 73 | 85 | 86 | 83 | 85 | Ligand 7 |
| I | 36 | 34 | 35 | 31 | 39 | 47 | 37 | 38 | 68 | 47 | 58 | 45 | 35 | 38 | 29 | 27 | 47 | 38 | 38 | 43 | 41 | 41 | 43 | 43 | Ligand 9 |
| J | 22 | 27 | 36 | 39 | 34 | 38 | 47 | 45 | 45 | 50 | 50 | 47 | 26 | 26 | 29 | 28 | 42 | 45 | 48 | 47 | 40 | 38 | 40 | 42 | Ligand 10 |
| K | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | Ligand 11 |
| L | 1 | 2 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Ligand 12 |
| M | 26 | 29 | 33 | 28 | 31 | 31 | 32 | 29 | 35 | 37 | 39 | 40 | 24 | 26 | 24 | 29 | 29 | 30 | 29 | 29 | 34 | 38 | 33 | 48 | Ligand 13 |
| N | 70 | 82 | 67 | 74 | 67 | 79 | 75 | 75 | 88 | 101 | 91 | 108 | 67 | 75 | 69 | 70 | 76 | 75 | 78 | 74 | 82 | 85 | 108 | 81 | Ligand 14 |
| O | 58 | 71 | 64 | 68 | 63 | 80 | 81 | 71 | 75 | 92 | 78 | 75 | 48 | 49 | 49 | 49 | 51 | 87 | 53 | 58 | 55 | 58 | 56 | 53 | Ligand 15 |
| P | 36 | 38 | 48 | 45 | 54 | 64 | 50 | 47 | 67 | 80 | 56 | 60 | 29 | 32 | 30 | 33 | 46 | 48 | 35 | 33 | 47 | 35 | 36 | 31 | Ligand 16 |
| | PhB(OH)$_2$ | | | | | | | | | | | | PhBpin | | | | | | | | | | | | |
| | Pd(OAc)$_2$ | | | | Pd(OTFA)$_2$ | | | | Pd(PhCN)$_2$Cl$_2$ | | | | Pd(OAc)$_2$ | | | | Pd(OTFA)$_2$ | | | | Pd(PhCN)$_2$Cl$_2$ | | | | |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 76 | 84 | 80 | 84 | 73 | 64 | 53 | 59 | 93 | 74 | 69 | 73 | 92 | 82 | 80 | 78 | 81 | 74 | 69 | 72 | 80 | 102 | 80 | 78 | Ligand 17 |
| B | 90 | 91 | 99 | 79 | 88 | 88 | 52 | 62 | 91 | 79 | 85 | 79 | 103 | 83 | 100 | 84 | 83 | 71 | 82 | 66 | 88 | 91 | 80 | 111 | Ligand 18 |
| C | 77 | 82 | 69 | 69 | 87 | 77 | 68 | 56 | 74 | 81 | 61 | 73 | 86 | 83 | 78 | 73 | 76 | 70 | 65 | 62 | 82 | 85 | 74 | 98 | Ligand 19 |
| D | 68 | 66 | 69 | 59 | 80 | 67 | 77 | 71 | 72 | 66 | 58 | 61 | 85 | 70 | 68 | 63 | 74 | 74 | 70 | 57 | 93 | 89 | 84 | 92 | Ligand 20 |
| E | 79 | 96 | 98 | 87 | 84 | 70 | 69 | 68 | 70 | 85 | 68 | 89 | 106 | 91 | 80 | 82 | 71 | 77 | 61 | 61 | 78 | 80 | 78 | 83 | Ligand 21 |
| F | 2 | 3 | 2 | 2 | 4 | 3 | 2 | 2 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 5 | 4 | 4 | 3 | 6 | 6 | 7 | 4 | Ligand 22 |
| G | 68 | 74 | 77 | 62 | 56 | 53 | 58 | 49 | 60 | 53 | 53 | 50 | 69 | 64 | 64 | 64 | 48 | 52 | 50 | 48 | 66 | 57 | 60 | 64 | Ligand 23 |
| H | 49 | 61 | 68 | 55 | 55 | 50 | 44 | 43 | 81 | 67 | 64 | 66 | 75 | 74 | 60 | 63 | 56 | 58 | 45 | 45 | 74 | 65 | 75 | 72 | Ligand 24 |
| I | 55 | 56 | 51 | 64 | 60 | 53 | 50 | 47 | 79 | 80 | 86 | 79 | 68 | 63 | 73 | 67 | 59 | 49 | 53 | 49 | 88 | 86 | 83 | 81 | Ligand 25 |
| J | 55 | 57 | 48 | 58 | 60 | 54 | 51 | 47 | 81 | 76 | 78 | 84 | 70 | 60 | 68 | 68 | 58 | 53 | 50 | 49 | 80 | 88 | 82 | 92 | Ligand 26 |
| K | 20 | 21 | 20 | 25 | 27 | 23 | 23 | 24 | 33 | 31 | 37 | 38 | 21 | 21 | 25 | 25 | 22 | 21 | 21 | 24 | 29 | 29 | 34 | 28 | Ligand 27 |
| L | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | Ligand 28 |
| M | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | Ligand 29 |
| N | 8 | 8 | 8 | 9 | 9 | 8 | 7 | 8 | 14 | 14 | 16 | 16 | 9 | 8 | 8 | 10 | 7 | 7 | 6 | 6 | 14 | 14 | 13 | 13 | Ligand 30 |
| O | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Ligand 31 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Ligand 32 |
| | PhB(OH)$_2$ | | | | | | | | | | | | PhBpin | | | | | | | | | | | | |
| | Pd(OAc)$_2$ | | | | Pd(OTFA)$_2$ | | | | Pd(PhCN)$_2$Cl$_2$ | | | | Pd(OAc)$_2$ | | | | Pd(OTFA)$_2$ | | | | Pd(PhCN)$_2$Cl$_2$ | | | | |

*Figure S5 Heatmap of the data obtained from HTE screening, lighter hues indicate higher conversion*

It should be noted that individual results that surpassed 100% yield were observed. Although the precise reason for this is extremely challenging to determine, analytical artifacts such as these can arise in the preparation of the analytical samples, the sampling of the analytical plate by the LC–MS autosampler, or during the analytical run. To mitigate the impact of such events on the overall data quality, each condition was run in quadruplicate and averaged. With each datapoint being the average of four repeats the impact that a single erroneous result can have on the overall dataset was thought to be minor and, as such, we decided not to eliminate any artificially high results (e.g., L7-C3) and instead treat all data points uniformly.

To provide a dataset for subsequent modelling, we calculated the average of all measurements which resulted in a dataset of 186 single datapoints. Whilst some single measurements had a yield above 100%, the averaged values did not.



| | Average Yields (%) Plate 1 | | | | | | Standard Deviation of Repeats Plate 1 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pd(OAc)$_2$ | Pd(OTFA)$_2$ | Pd(PhCN)$_2$Cl$_2$ | Pd(OAc)$_2$ | Pd(OTFA)$_2$ | Pd(PhCN)$_2$Cl$_2$ | | | | | | |
| Ligand 0 | 1 | 2 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 1 |
| Ligand 1 | 18 | 31 | 24 | 14 | 24 | 20 | 1 | 5 | 1 | 3 | 1 | 1 |
| Ligand 2 | 22 | 35 | 35 | 17 | 26 | 20 | 3 | 3 | 1 | 1 | 3 | 3 |
| Ligand 3 | 87 | 73 | 98 | 75 | 73 | 89 | 2 | 6 | 2 | 4 | 7 | 6 |
| Ligand 4 | 51 | 65 | 72 | 43 | 77 | 59 | 7 | 9 | 7 | 5 | 3 | 3 |
| Ligand 5 | 78 | 77 | 100 | 61 | 91 | 94 | 6 | 12 | 4 | 2 | 10 | 13 |
| Ligand 6 | 39 | 41 | 59 | 32 | 43 | 45 | 3 | 3 | 7 | 1 | 3 | 1 |
| Ligand 7 | 70 | 59 | 102 | 56 | 72 | 85 | 6 | 4 | 11 | 1 | 5 | 1 |
| Ligand 9 | 34 | 40 | 55 | 32 | 41 | 42 | 2 | 5 | 11 | 5 | 4 | 1 |
| Ligand 10 | 31 | 41 | 48 | 27 | 45 | 40 | 8 | 6 | 3 | 2 | 3 | 2 |
| Ligand 11 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ligand 12 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Ligand 13 | 29 | 31 | 38 | 26 | 29 | 38 | 3 | 1 | 2 | 2 | 1 | 7 |
| Ligand 14 | 73 | 74 | 97 | 70 | 76 | 89 | 7 | 5 | 9 | 3 | 2 | 13 |
| Ligand 15 | 65 | 74 | 80 | 49 | 62 | 56 | 6 | 8 | 8 | 1 | 17 | 2 |
| Ligand 16 | 42 | 54 | 66 | 31 | 41 | 37 | 5 | 7 | 11 | 2 | 8 | 7 |

| Standard Deviation of Repeats Plate 2 | | | | | | Average Yields (%) Plate 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Pd(OAc)$_2$ | Pd(OTFA)$_2$ | Pd(PhCN)$_2$Cl$_2$ | Pd(OAc)$_2$ | Pd(OTFA)$_2$ | Pd(PhCN)$_2$Cl$_2$ | |
| 4 | 8 | 11 | 6 | 5 | 11 | 81 | 62 | 77 | 83 | 74 | 85 | Ligand 17 |
| 8 | 18 | 6 | 10 | 8 | 13 | 90 | 73 | 83 | 93 | 75 | 93 | Ligand 18 |
| 6 | 13 | 8 | 6 | 6 | 10 | 74 | 72 | 73 | 80 | 68 | 84 | Ligand 19 |
| 4 | 6 | 6 | 9 | 8 | 4 | 66 | 74 | 64 | 71 | 69 | 89 | Ligand 20 |
| 9 | 8 | 11 | 12 | 8 | 3 | 90 | 73 | 78 | 90 | 67 | 80 | Ligand 21 |
| 0 | 1 | 0 | 0 | 1 | 1 | 2 | 3 | 3 | 3 | 4 | 6 | Ligand 22 |
| 7 | 4 | 4 | 3 | 2 | 4 | 70 | 54 | 54 | 65 | 49 | 62 | Ligand 23 |
| 8 | 6 | 8 | 8 | 7 | 5 | 59 | 48 | 70 | 68 | 51 | 72 | Ligand 24 |
| 6 | 6 | 3 | 4 | 5 | 3 | 57 | 53 | 81 | 68 | 52 | 84 | Ligand 25 |
| 4 | 5 | 4 | 5 | 4 | 5 | 54 | 53 | 80 | 66 | 52 | 86 | Ligand 26 |
| 2 | 2 | 3 | 2 | 1 | 3 | 21 | 24 | 35 | 23 | 22 | 30 | Ligand 27 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Ligand 28 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Ligand 29 |
| 1 | 1 | 1 | 1 | 0 | 1 | 8 | 8 | 15 | 8 | 6 | 14 | Ligand 30 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | Ligand 31 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | Ligand 32 |

*Figure S6 Condensed heatmap alongside the standard deviation of repeated measurements*

## Generation of Morgan Fingerprints

All reagent molecules (pre-catalyst, ligands, boronates) were drawn using ChemDraw and then the SMILES were generated. The SMILES were canonicalized using the package RDKit in python. Then RDKit (RDKit version 2020.03.2) was used to generate Morgan fingerprints of a radius 2 with a set length of 1024. The three fingerprints were concatenated and saved as Numpy arrays for subsequent modelling. Initially, different radii of circular fingerprints of the ligand molecules were screened, and it was observed that a radius of 2 was ideal due to the lowest model error. Figure S7 illustrates different radii of Morgan fingerprints versus their RMSE of three different ML models (RF, GP, ANN) – the RMSE is averaged from three single evaluation (prediction of yield) of a random split of the data (80% training, 20% test data) and the error bars represent the standard deviation.

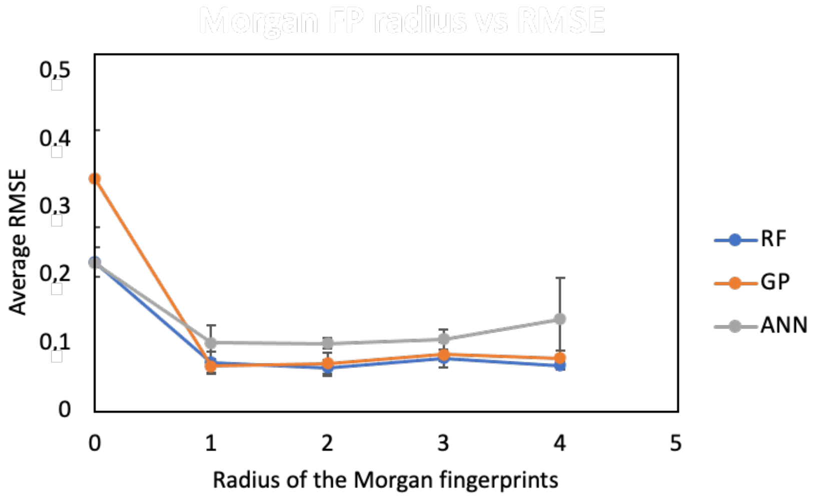


*Figure S7 Variation of the used radius for fingerprint generation. The presented values are averaged from three single evaluations using random split (80/20 : train/test) and the error bars represent the standard deviation.*

Based on the generated fingerprints we conducted a similarity assessment between all ligands using the Tanimoto similarity index (using RDKit) (see heatmap in Figure S8a). Moreover, hierarchical clustering was conducted (using the SciPy python package) to allow for insights into similarities between the ligands and understand which are structurally close.

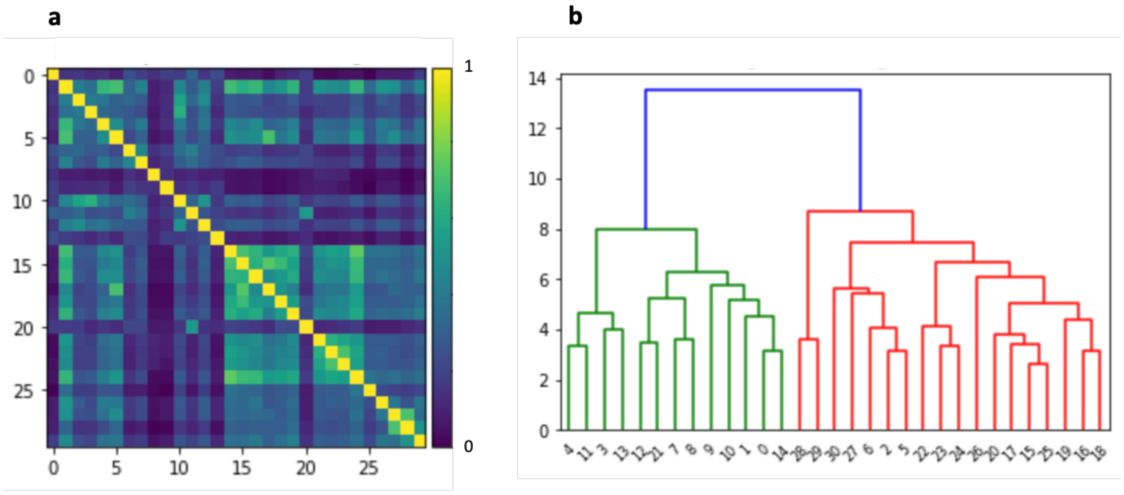


*Figure S8 Similarity assessment of the ligand scope (a) Tanimoto similarity index between all ligands (b) Dendrogram showing hierarchical clustering of all ligands*

## Density Functional Theory (DFT)-based Geometry Optimization

For DFT geometry optimizations we relied on the B3LYP functional and 6-31G(d) basis set (Gaussian 16). As stated in the manuscript, DFT was used for unbound ligand molecules only. This section details the generation of steric and electronic descriptors. In addition to the features explained below we also calculated HOMO/LUMO energies of the ligand molecules. Table S3 displays all DFT derived descriptor values.

### Sterimol Parameters

Sterimol descriptors comprise of three single length measurements that capture the steric footprint of a molecule across a specified axis and relative to a fixed point of reference. All calculations were conducted using a python package developed by Brethomé *et al.* and the geometry optimized ligand molecules.[2] Parameterization was separated into the $\alpha$-carbon residue (Figure S9, R-res) and the acetyl residue (Figure S9, N-res). The arrows in the figure indicate the direction of the reference axis.

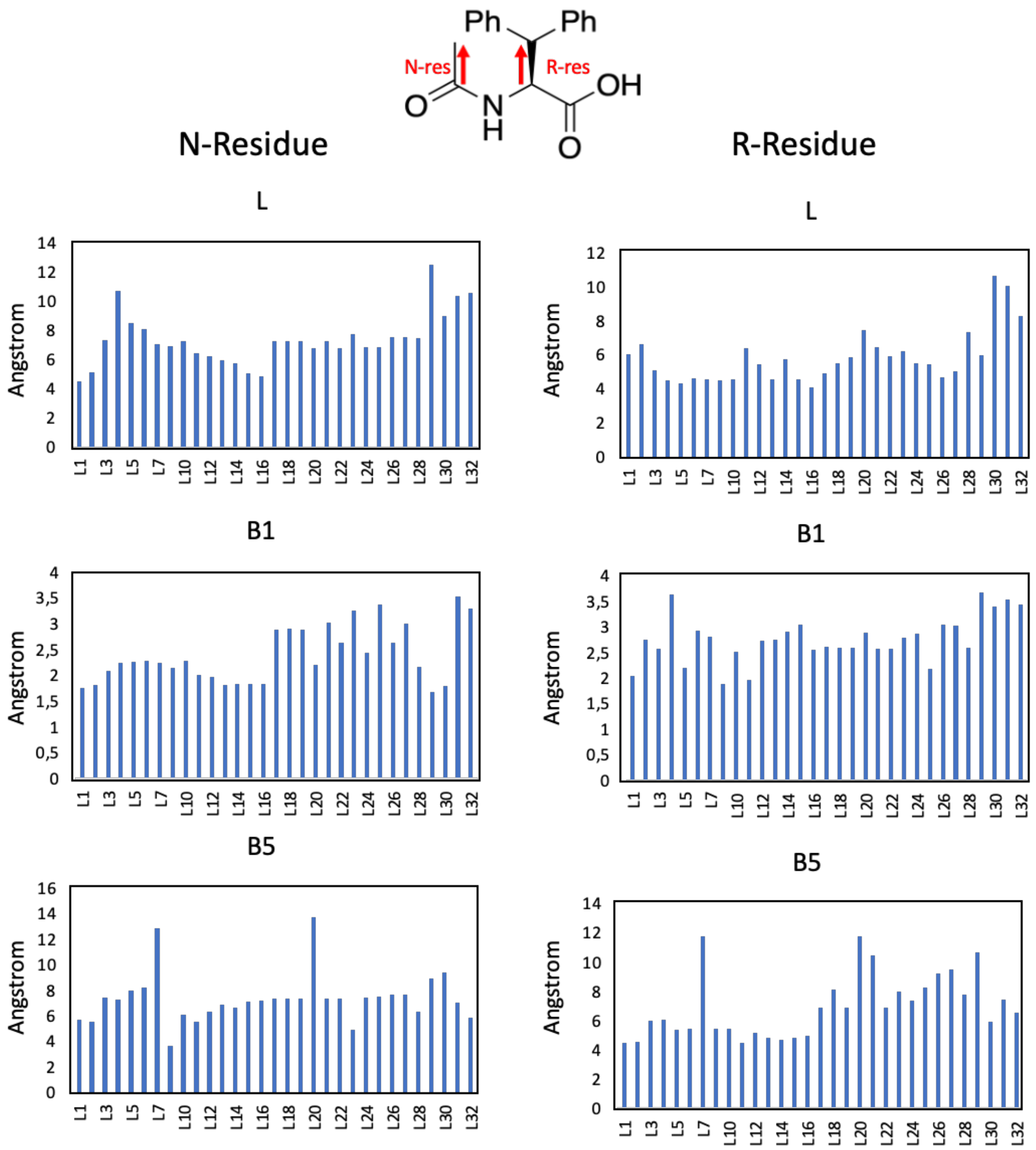


*Figure S9 Sterimol parameters of the N-residue and the R-residue. The molecule on top indicates the reference axis for generation of the Sterimol parameters and the six plots show the results of the calculation.*

## Percentage Buried Volume

The bulkiness of the $\alpha$-carbon residue was additionally quantified by calculating the percentage of the buried volume, based on geometry optimized ligand molecules. In this case the $\alpha$-carbon was set as center and the calculation was conducted regarding solely the residue on the $\alpha$-carbon position, using the SambVca 2.1 web application[3] as shown in Figure S10. The used reference pane of the 4 neighbouring atoms and the direction of the reference axis is shown in Figure S10a. Figure S10b illustrates the part of the molecules (R-res) that was considered for the calculation (here a *tert*-butyl group) and Figure S10c is a two-dimensional steric heatmap of the outcome of the calculation of the % buried volume. The graph (Figure S10d) displays the results of the calculation and allows for steric insights into the bulkiness of the ligand molecule.

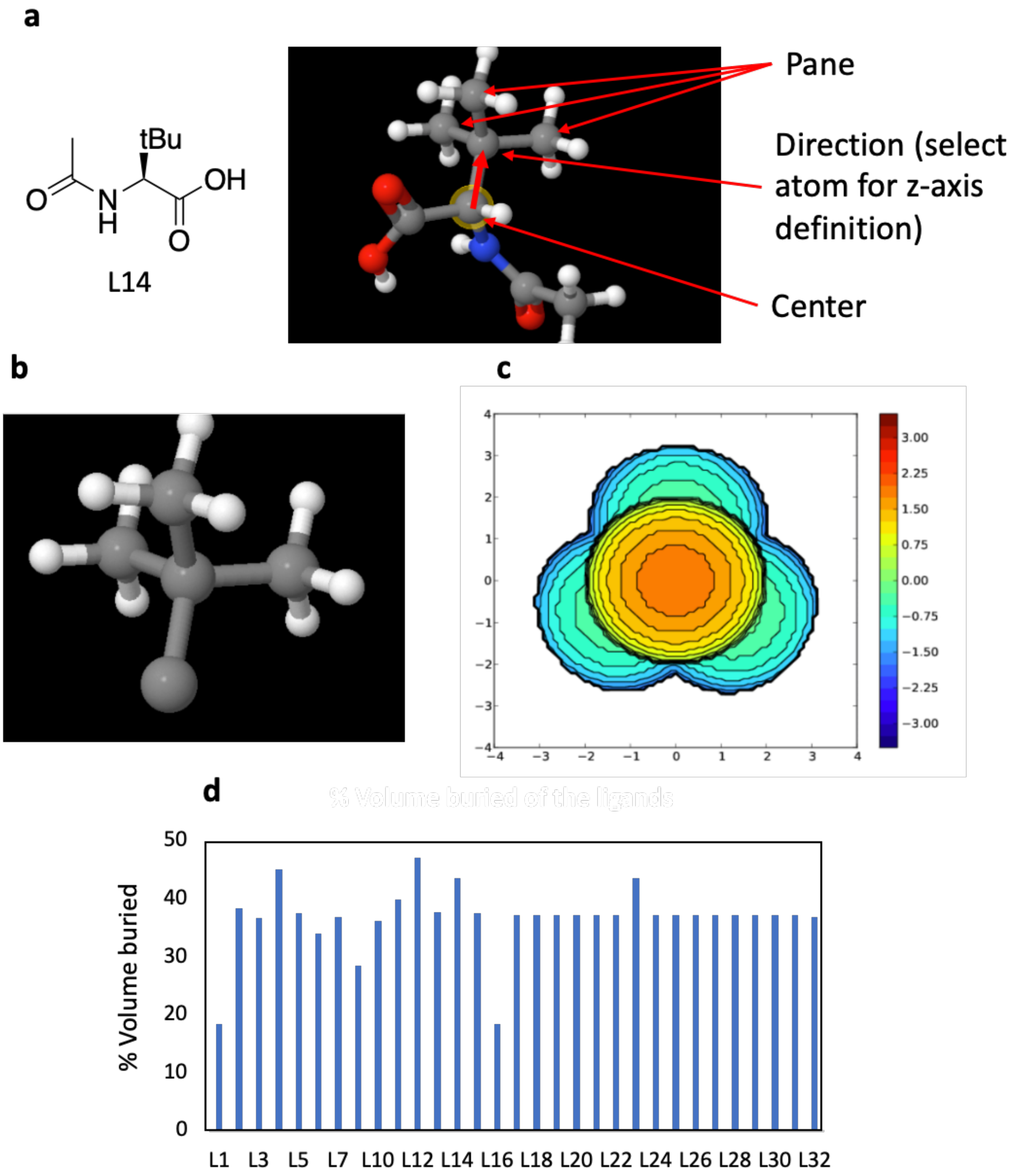


*Figure S10 Calculation of the percentage buried volume (a) Structure and DFT optimized geometry of ligand 14. Illustration of the center, reference axis and pane for calculation of the % buried volume (b) Illustration of the actual R-residue (here tert-butyl) which was considered for the calculation (c) Graphic illustration of the two-dimensional steric heatmap of the calculation of the % buried volume from the web-based platform (d) Results of the % buried volume calculation.*

## Natural Bond Orbital (NBO) Analysis

In order to capture the electron density distribution, we conducted a NBO analysis using Gaussian 16.[4] Figure S11 illustrates the location of the atoms in the ligand molecule which were selected for NBO analysis as well as the results of the calculation.

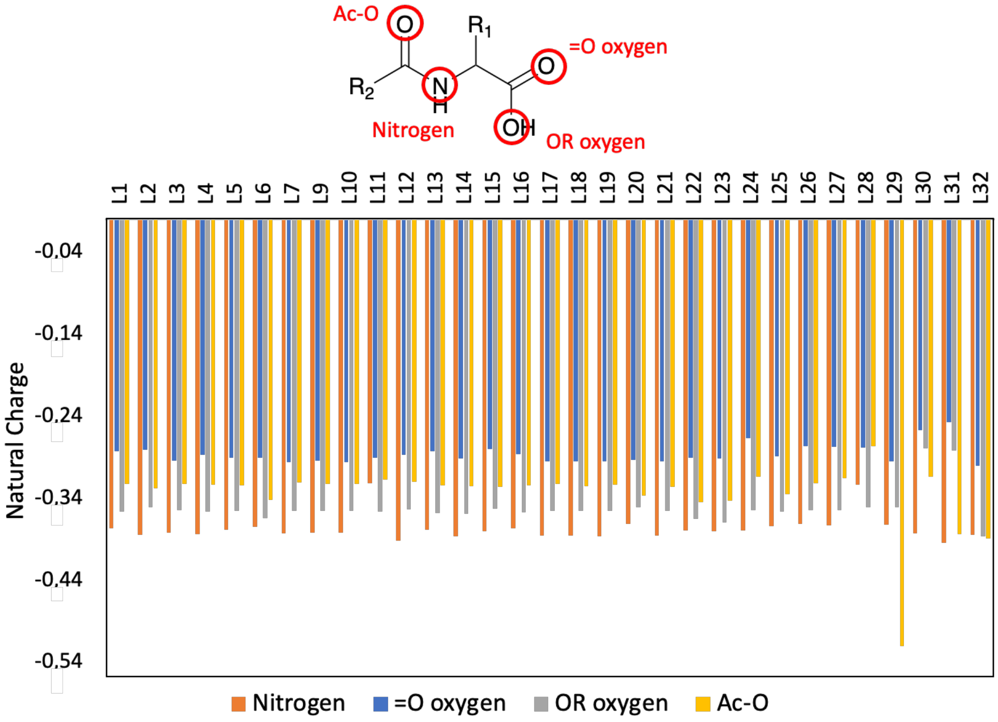


*Figure S11 Location of the atoms which were used for the NBO analysis and results of the NBO calculation of all ligands.*

## CHarges from ELectrostatic Potentials Using a Grid-Based Method (ChELPG) Analysis

CHELPG[5] analysis was conducted for all ligands using Gaussian 16. Figure S12 illustrates the location of the atoms in the ligand molecule which were selected for CHELPG analysis as well as the results of the calculation.

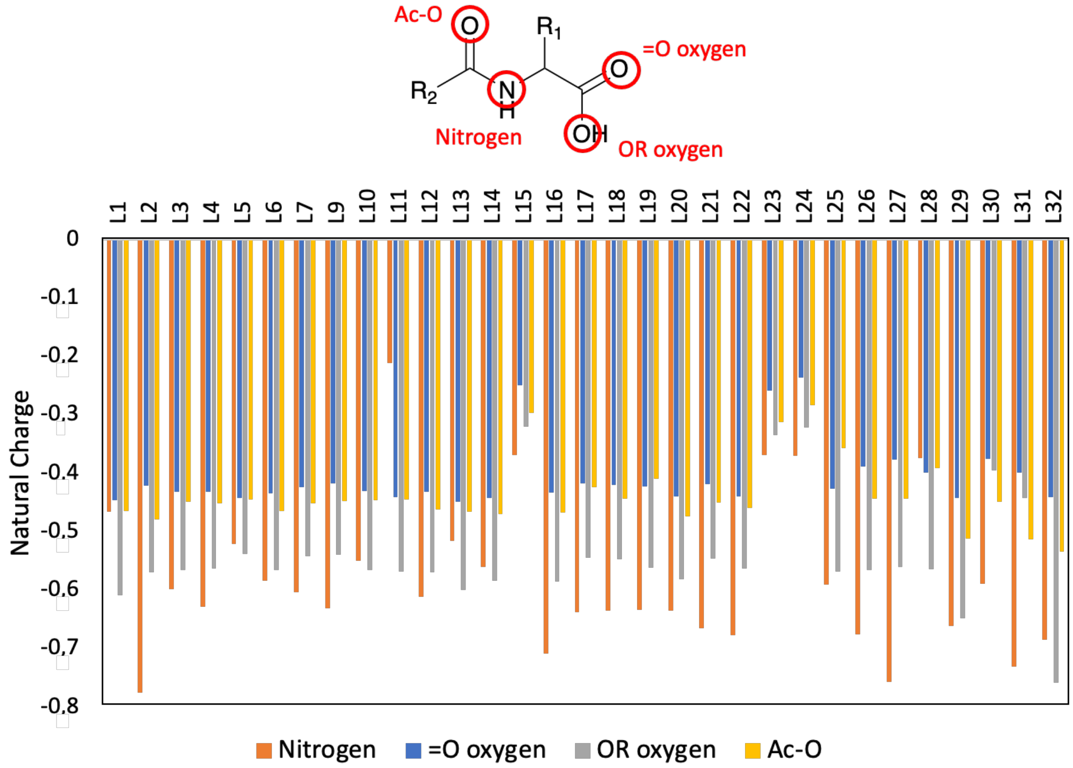


*Figure S12 Location of the atoms which were used for the CHELPG analysis and results of the CHELPG calculation of all ligands.*

## Summary of DFT Descriptor Values

Table S3 includes all values of the DFT based descriptors.

*Table S3 Summary of all calculated DFT descriptors*

| Ligands | Sterimol R-residue | | | Sterimol N-residue | | | NBO analysis | | | | CHELPG analysis | | | | % buried volume |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L | B1 | B5 | L | B1 | B5 | N | OR | =O | Ac-O | N | OR | =O | Ac-O | |
| L1 | 6.06 | 2.05 | 4.51 | 4.54 | 1.76 | 5.73 | -0.37619 | -0.35586 | -0.28253 | -0.3226 | -0.466513 | -0.608944 | -0.446147 | -0.464045 | 18.4 |
| L2 | 6.66 | 2.75 | 4.56 | 5.12 | 1.82 | 5.56 | -0.38442 | -0.35069 | -0.28116 | -0.32745 | -0.775537 | -0.569357 | -0.421585 | -0.478563 | 38.4 |
| L3 | 5.1 | 2.58 | 6.03 | 7.36 | 2.09 | 7.43 | -0.38195 | -0.35478 | -0.29432 | -0.32243 | -0.599068 | -0.565653 | -0.432344 | -0.448455 | 36.7 |
| L4 | 4.49 | 3.64 | 6.08 | 10.74 | 2.26 | 7.29 | -0.38321 | -0.35592 | -0.28733 | -0.32306 | -0.629191 | -0.563101 | -0.431461 | -0.451794 | 45.1 |
| L5 | 4.36 | 2.21 | 5.37 | 8.54 | 2.27 | 7.96 | -0.37789 | -0.35516 | -0.29067 | -0.32381 | -0.520643 | -0.538302 | -0.441716 | -0.444796 | 37.6 |
| L6 | 4.62 | 2.93 | 5.47 | 8.08 | 2.29 | 8.23 | -0.37512 | -0.36435 | -0.29063 | -0.34228 | -0.583526 | -0.565598 | -0.433752 | -0.464669 | 34 |
| L7 | 4.58 | 2.81 | 11.79 | 7.05 | 2.25 | 12.85 | -0.38234 | -0.35532 | -0.29576 | -0.32062 | -0.603403 | -0.542356 | -0.423286 | -0.451632 | 36.9 |
| L9 | 4.49 | 1.9 | 5.45 | 6.94 | 2.15 | 3.7 | -0.38201 | -0.35491 | -0.29448 | -0.32238 | -0.630682 | -0.538847 | -0.417083 | -0.448027 | 28.5 |
| L10 | 4.6 | 2.52 | 5.43 | 7.3 | 2.29 | 6.1 | -0.3818 | -0.35569 | -0.29587 | -0.32223 | -0.549897 | -0.566176 | -0.430958 | -0.446708 | 36.3 |
| L11 | 6.4 | 1.96 | 4.51 | 6.45 | 2.02 | 5.55 | -0.32148 | -0.35629 | -0.29017 | -0.31724 | -0.211833 | -0.56818 | -0.440953 | -0.444532 | 40 |
| L12 | 5.49 | 2.74 | 5.21 | 6.28 | 1.98 | 6.31 | -0.39196 | -0.35362 | -0.28726 | -0.3199 | -0.61206 | -0.569427 | -0.4315 | -0.462546 | 47.2 |
| L13 | 4.55 | 2.76 | 4.87 | 5.96 | 1.83 | 6.92 | -0.37844 | -0.35757 | -0.28296 | -0.32388 | -0.515195 | -0.599674 | -0.448818 | -0.465924 | 37.7 |
| L14 | 5.77 | 2.92 | 4.67 | 5.76 | 1.85 | 6.66 | -0.38598 | -0.35873 | -0.29181 | -0.32543 | -0.560531 | -0.584289 | -0.442257 | -0.470392 | 43.7 |
| L15 | 4.56 | 3.04 | 4.86 | 5.07 | 1.84 | 7.11 | -0.37986 | -0.35286 | -0.27977 | -0.32606 | -0.368372 | -0.320297 | -0.248912 | -0.296172 | 37.6 |
| L16 | 4.12 | 2.56 | 5.01 | 4.84 | 1.84 | 7.23 | -0.37648 | -0.35736 | -0.28612 | -0.32406 | -0.708605 | -0.585613 | -0.433567 | -0.467369 | 18.4 |
| L17 | 4.94 | 2.61 | 6.89 | 7.27 | 2.9 | 7.4 | -0.38537 | -0.35497 | -0.29515 | -0.32284 | -0.637421 | -0.544184 | -0.417856 | -0.424352 | 37.3 |
| L18 | 5.53 | 2.6 | 8.17 | 7.27 | 2.91 | 7.37 | -0.38569 | -0.35496 | -0.29525 | -0.32549 | -0.635056 | -0.546674 | -0.420481 | -0.443747 | 37.3 |
| L19 | 5.86 | 2.6 | 6.93 | 7.29 | 2.9 | 7.38 | -0.38591 | -0.35521 | -0.29524 | -0.32314 | -0.633977 | -0.561779 | -0.422134 | -0.409802 | 37.3 |
| L20 | 7.46 | 2.89 | 11.81 | 6.77 | 2.22 | 13.77 | -0.37146 | -0.35119 | -0.29294 | -0.33671 | -0.635565 | -0.581287 | -0.440082 | -0.474375 | 37.2 |
| L21 | 6.47 | 2.58 | 10.51 | 7.27 | 3.04 | 7.35 | -0.38579 | -0.35495 | -0.29528 | -0.3262 | -0.665955 | -0.545958 | -0.418825 | -0.450242 | 37.3 |
| L22 | 5.95 | 2.57 | 6.93 | 6.81 | 2.64 | 7.34 | -0.37901 | -0.36471 | -0.29035 | -0.34504 | -0.676505 | -0.562611 | -0.440287 | -0.458779 | 37.3 |
| L23 | 6.25 | 2.79 | 7.98 | 7.76 | 3.26 | 4.92 | -0.3798 | -0.36903 | -0.29149 | -0.34273 | -0.369425 | -0.334984 | -0.259106 | -0.312619 | 43.6 |
| L24 | 5.51 | 2.87 | 7.4 | 6.9 | 2.44 | 7.44 | -0.37897 | -0.35457 | -0.26648 | -0.3135 | -0.370268 | -0.322206 | -0.236543 | -0.283715 | 37.3 |
| L25 | 5.47 | 2.18 | 8.31 | 6.9 | 3.38 | 7.49 | -0.37417 | -0.3565 | -0.28921 | -0.33472 | -0.590294 | -0.568786 | -0.426612 | -0.35735 | 37.3 |
| L26 | 4.67 | 3.04 | 9.24 | 7.57 | 2.65 | 7.66 | -0.37087 | -0.3543 | -0.27604 | -0.32191 | -0.67638 | -0.565751 | -0.388932 | -0.443194 | 37.3 |
| L27 | 5.06 | 3.03 | 9.55 | 7.56 | 3.02 | 7.7 | -0.373 | -0.35399 | -0.27696 | -0.31547 | -0.756809 | -0.560124 | -0.377225 | -0.44324 | 37.3 |
| L28 | 7.32 | 2.59 | 7.83 | 7.48 | 2.17 | 6.34 | -0.32331 | -0.35088 | -0.27854 | -0.27649 | -0.374424 | -0.564444 | -0.398645 | -0.391023 | 37.3 |
| L29 | 6 | 3.68 | 10.69 | 12.5 | 1.68 | 8.97 | -0.372 | -0.35088 | -0.29467 | -0.5198 | -0.660768 | -0.647813 | -0.442392 | -0.511759 | 37.3 |
| L30 | 10.62 | 3.41 | 5.97 | 8.97 | 1.8 | 9.38 | -0.38254 | -0.27938 | -0.25713 | -0.31348 | -0.588752 | -0.395623 | -0.375246 | -0.448494 | 37.3 |
| L31 | 10.04 | 3.54 | 7.48 | 10.4 | 3.54 | 7.03 | -0.39434 | -0.28205 | -0.24733 | -0.38337 | -0.730678 | -0.44217 | -0.398695 | -0.513599 | 37.3 |
| L32 | 8.3 | 3.45 | 6.56 | 10.6 | 3.31 | 5.84 | -0.38476 | -0.38625 | -0.30004 | -0.38919 | -0.684506 | -0.758856 | -0.441578 | -0.533879 | 36.9 |

## Machine Learning

Within this study five different ML surrogate models were used. Information on the chosen hyperparameters, the used software packages and the implementation can be found in this section. ML hyperparameter tuning was conducted for all models using feature set 14 within the initial supervised ML (random split). We observed that tuning hyperparameters separately for each feature set did not deliver significantly better performance than using the hyperparameters obtained when using feature set 14.

Figure S13 illustrates the workflow of parameterizing inputs and subsequentially conducting initial supervised ML and active ML featured closed-loop optimization.

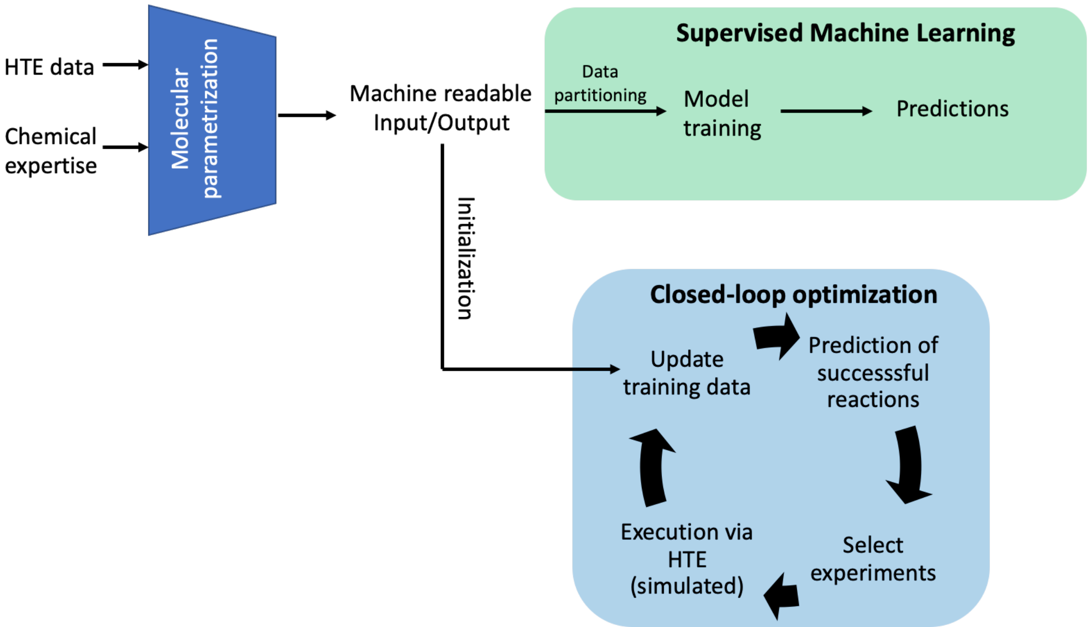

HTE data
Chemical expertise
Molecular parametrization
Machine readable Input/Output
Supervised Machine Learning
Data partitioning
Model training
Predictions
Initialization
Closed-loop optimization
Update training data
Prediction of successsful reactions
Select experiments
Execution via HTE (simulated)

*Figure S13 ML workflow in this study: The input data was parameterized and then used for supervised ML studies and for active ML.*

### Linear Model

We implemented the linear model using the package scikit-learn, version 0.23.0 – this version was used for all subsequent modelling. Unless explicitly stated, all parameters were left at default values.

### Random Forest

We implemented the random forest surrogate model using the package scikit-learn. A total number of 200 estimators were used consistently for modelling. Unless stated otherwise all parameters were kept to default. Unless explicitly stated, all parameters were left at default values.

### Gaussian Process

We implemented the Gaussian process model using the package scikit-learn. As covariance function we used a Matèrn 3/2 kernel with a common length scale for all inputs. We observed that adding a white kernel to account for measurement noise did not improve the prediction performance and was hence omitted.

```
kernel = 1.0 * Matern(length_scale=1.2, nu=1.5)
```

### Artificial Neural Network

The artificial neural network model was implemented using the Tensorflow Keras 2.3.0. The fully connected feed-forward network consisted of six hidden layers of ten nodes each and ReLu activation function. The final layer consisted of one single node. The weights and biases were initialized with the default schemes (Glorot uniform and zeros, respectively). Training was done with RMSProp using default parameters over 1000 epochs with a minibatch size of 32.

### Adaptive Boosting Model

We implemented the AdaBoost model using the package scikit-learn. A total number of 200 estimators were used consistently for all modelling.

### Support Vector Regression

We implemented the support vector regression model *via* the package scikit-learn using a linear kernel.

## Leave-one-group-out (LOGO) Cross Validation (CV)

LOGO CV was used within the study of supervised ML to assess the models' performance to conduct extrapolative predictions. Figure S14 allows for insights into the single folds of the LOGO CV using feature set 14. This plot illustrates the test/train RMSE (y-axis) of the single groups (x-axis, 1-31), highlighting that the modelling performance strongly varies from ligand to ligand. Additionally, it is visible how different models fit train/test data.

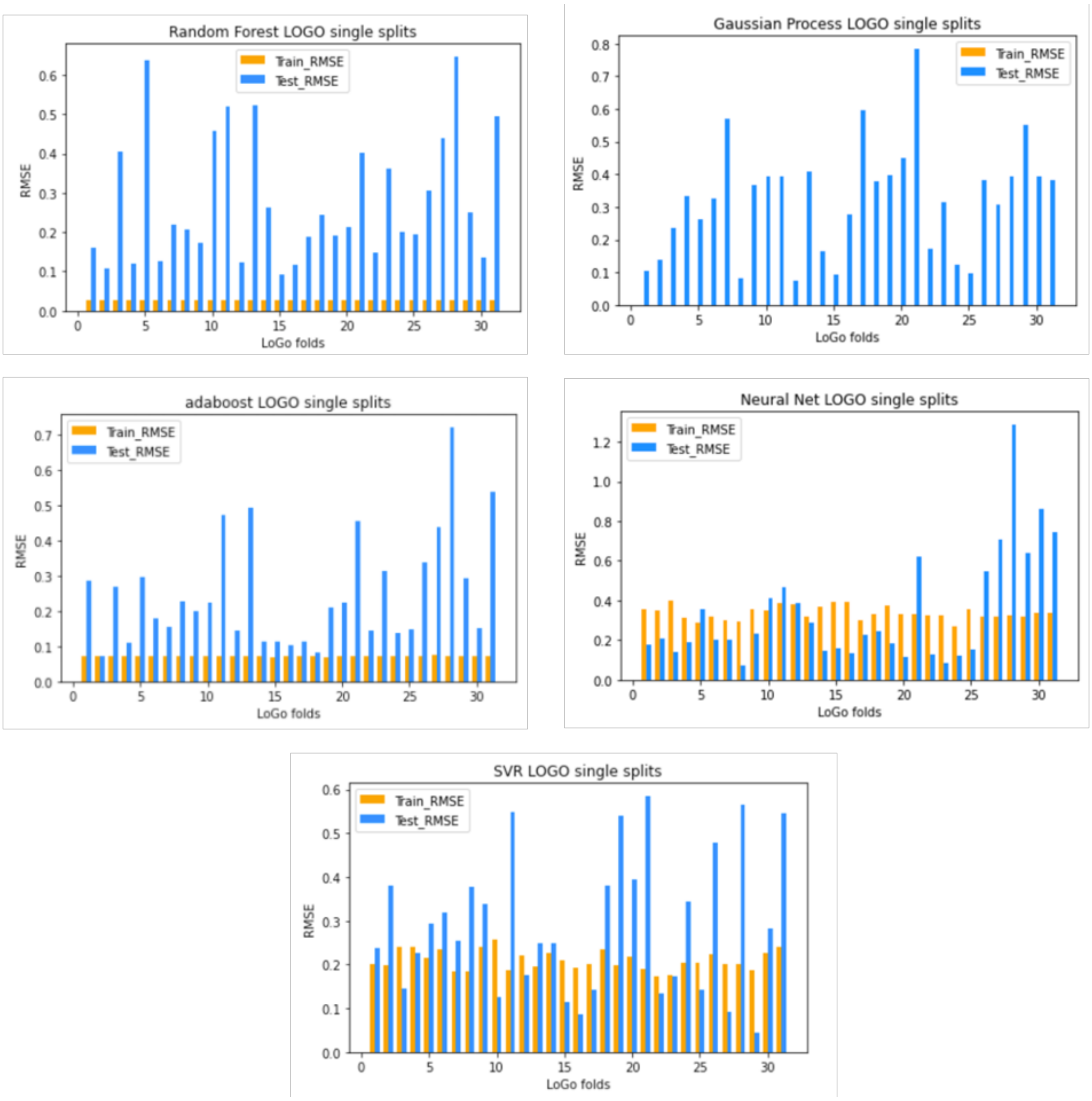


*Figure S14 Detailed insights into the LOGO CV of different surrogate models*

To investigate a potential correlation of the variation of the different train/test folds on model performance we looked at structural similarity. Using the Tanimoto similarity index,[6] the similarity between the training data (30 of the 31 ligands) and the test data (1 of the 31 ligands) was calculated and then the average of the values was taken for all 31 ligand molecules. Figure S15 shows the relationship between the averaged similarity indices and the RMSE of all 31 folds. When overlaying the results from all 5 models a higher density of datapoints in the lower right quadrant of the plot suggests that higher similarity between test and train data delivers lower model error, as expected. This follows the rational considerations of ML that model performance is typically increased when the training and test data have a higher similarity.

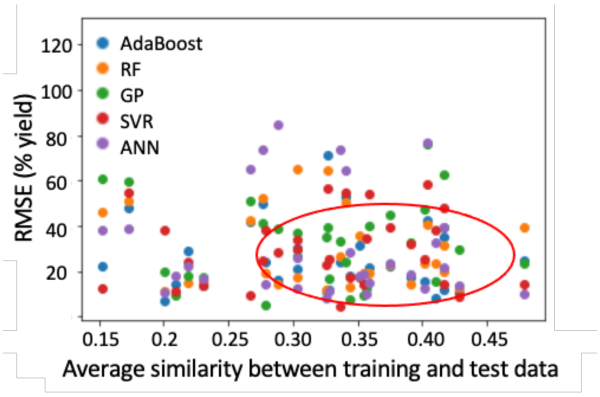
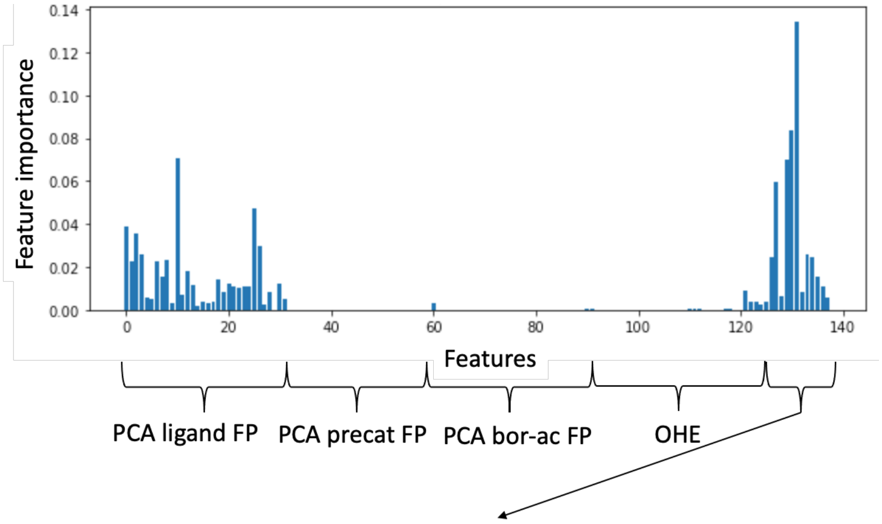


*Figure S15 Insights into the LOGO CV evaluation - RMSE of the fold vs train-test data similarity using Tanimoto similarity index (feature set 14). The circle indicates increased density of datapoints, suggesting that a higher similarity leads to a lower RMSE. This follows general considerations of ML that increased similarity of training and test data delivers better performance.*

## Feature Importance Assessment of the Random Forest

Explainable AI (XAI) is a field which attempts to convert a 'black-box' model into a 'white-box' model throught increased transparency, ultimately allowing for an explanations/justifications as to how certain predictions were made.[7, 8] The application of XAI in synthetic chemistry challenges should therefore allow for increased understanding of the chemical system and enable synthetic chemists to profit from the pattern recognition-based strengths possessed by ML. While black-box models often keep these patterns hidden within the model architecture, highlighting this information can deliver great benefits. Within this project we attempted XAI with a feature importance assessment – here a RF model which was trained on the complete hybrid feature set was subsequently analyzed using Gini importance.[9] Figure S16 illustrates the feature importance, highlighting that the PCA of the fingerprints of the ligands contain relevant information while those of the pre-catalyst/boronic acid seem redundant. Overall, OHE also has a very low importance when combined with the other features in feature set 14.



Feature: 126, Score: 0.02462 → HOMO
Feature: 127, Score: 0.05956 → LUMO
Feature: 128, Score: 0.00629 → NBO_N
Feature: 129, Score: 0.07011 → NBO_OR
Feature: 130, Score: 0.08362 → NBO_O__
Feature: 131, Score: 0.13443 → NBO_AcO
Feature: 132, Score: 0.00859 → chelpg_N
Feature: 133, Score: 0.02560 → chelpg_OR
Feature: 134, Score: 0.02451 → chelpg_O
Feature: 135, Score: 0.01562 → chelpg_AcO
Feature: 136, Score: 0.01073 → vol_bur
Feature: 137, Score: 0.00575 → n_proton

*Figure S16 Variation of the feature importance for RF models and detailed insights into relevant features of the hybrid feature set 14*

It must be noted that the feature importance slightly changes every time a RF is retrained as the single trees within the forest are populated differently. Nonetheless, the highest importance was predicted to be the NBO analysis of the amide oxygen, thus suggesting that the electron density around the oxygen matters. This observation reflects the known elements of the concerted metalation deprotonation (CMD) mechanism, with the acetamide oxygen functioning as an internal base that abstracts the proton. This demonstrates the ability of XAI to deliver chemical insights on complex systems. Moreover, the insignificance of the features encoding for the pre-catalyst and the boronic acid aligns with the fact that those parameters do have a high impact on reaction yield. As visible in Figure S5, the reaction outcome is mainly influenced by the ligands, rather than the identity of the pre-catalysts or boronates. This is clearly visible as in several rows of the heatmap (row = ligand) the whole row is either dark blue (low yield) or green/yellow (high yield), however, within the columns of the line there is only limited variation in yield.

# Closed-loop Optimization

## Expected Improvement Acquisition Function

Compared to a purely exploitative search within the closed-loop optimization, using the EI acquisition function allows for a controlled trade-off between exploitation and exploration. Any parameter combination $\theta$ delivers a predicted mean $\theta$(μ) and a standard deviation $\sigma(\theta)$. Following Eqn. 6, EI can be calculated relatively with respect to the current best condition from previous iterations, referred to as $m_{opt}$.

$$EI\left(\theta;\, m_{opt}\right) = \delta(\theta)\Phi\left(\frac{\delta(\theta)}{\sigma(\theta)}\right) + \sigma(\theta)\phi\left(\frac{\delta(\theta)}{\sigma(\theta)}\right) \qquad \text{Eqn 6.}$$

where $\delta(\theta) = \mu(\theta) - m_{opt}$, $\Phi$ is cumulative standard normal,

$\phi$ is standard normal density

The distance to the best condition is calculated by $\delta(\theta)$ and the search is conducted with the objective to find the $\theta$ that maximizes EI. It is noteworthy that not all ML models deliver an uncertainty metric, for example, GPs have built-in variance due to the model design, whereas ANN and RF do not have an uncertainty output. It should be noted that in our study we navigate in a solely discrete optimization space.

## De-full Factorization of the Chemical Space Study

We hypothesized that the simplicity of OHE along with a full factorial space could be more beneficial when compared to the effect on other input features (e.g. hybrid inputs) which are far more complex and might represent a challenge for the model to detect patterns in the data. To test this assumption, we dropped a random selection of the datapoints of the entire dataset (25%), therefore no longer representing a full factorial chemical space. However, we still observed that OHE outperformed the full feature set (Figure S17a). Figure S17b illustrates that even though the dataset was reduced, yield was still well distributed.

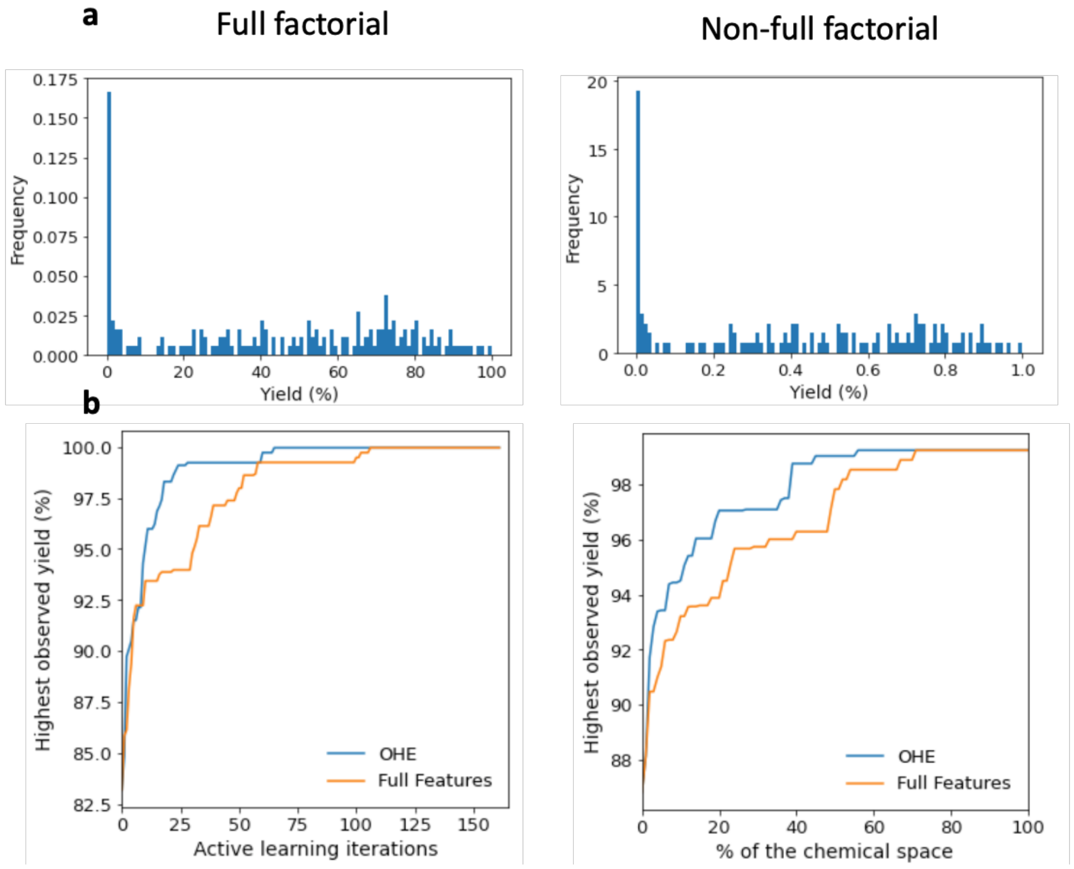


*Figure S17 Comparison of active learning model performance of a full factorial and a non-full factorial chemical space (a) Yield distribution of full factorial and a non-full factorial chemical space (b) Active learning curves of full factorial and a non-full factorial chemical space*

## Batch-Sequential Active Learning

We assessed the impact of using different batch sizes *vs* using sequential sampling. In Figure S 18 (x-axis normalized) the batch size was varied between two and 25 experiments (during each iteration) and sequential sampling (one experiment at a time) is illustrated as a baseline. The minimal differences in the learning curves indicate that smaller batch sizes have favorable learning curves compared to larger batch sizes. We hypothesize that a smaller batch size allows the active ML model to be updated more frequently and thus conduct predictions of slightly higher accuracy. At 40% of the chemical space (Figure 18, x-axis) the active learning strategy using a batch size of 25 iterated two times whereas the batch size of two iterated 30 times. Overall though, it seems that the batch size does not significantly impact the learning trajectory and thus the size should mainly be chosen based on experimental restrictions (e.g. possible number of experiments which can be run in parallel).

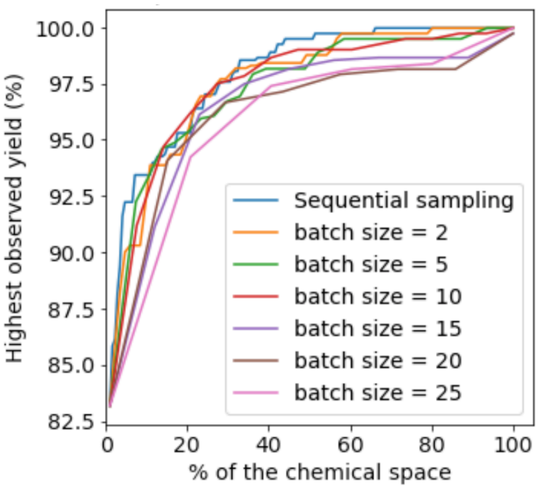


*Figure S18 Comparison of different batch sizes for active learning using RF and feature set 14*

## The Impact of Initialization of the Active Learning

The efficiency of closed-loop optimization algorithms depends on the data on which the very initial model is trained. Herein, we are comparing a broader set of reaction conditions (on average the dataset contains information of 7 ligands) to a restricted dataset (the dataset contains datapoints of only 3 ligands). To allow for general statements, the ligand in the train/test set were varied during 10 single experiments and the average of the learning curves was used for the plot. As visible in Figure S19, the location of the initialization data, and the lookup table (of one experiment) were illustrated in a dimensionality reduced 2D map that was generated using the first two principal components of the Morgan 2 fingerprints of all components. Whilst Figure S19a illustrates the initial data being randomly distributed over the chemical space, Figure S19b has the initial data located close to each other such that it has been intentionally limited to only to three ligands. It must be noted that the plot contains an overlap of datapoints which is a result of the dimensionality reduction. In Figure S19c, it is apparent that even as the local initialization possessed restricted knowledge, within ten iterations the model performance was approximately equal to an initialization dataset which is more diverse.

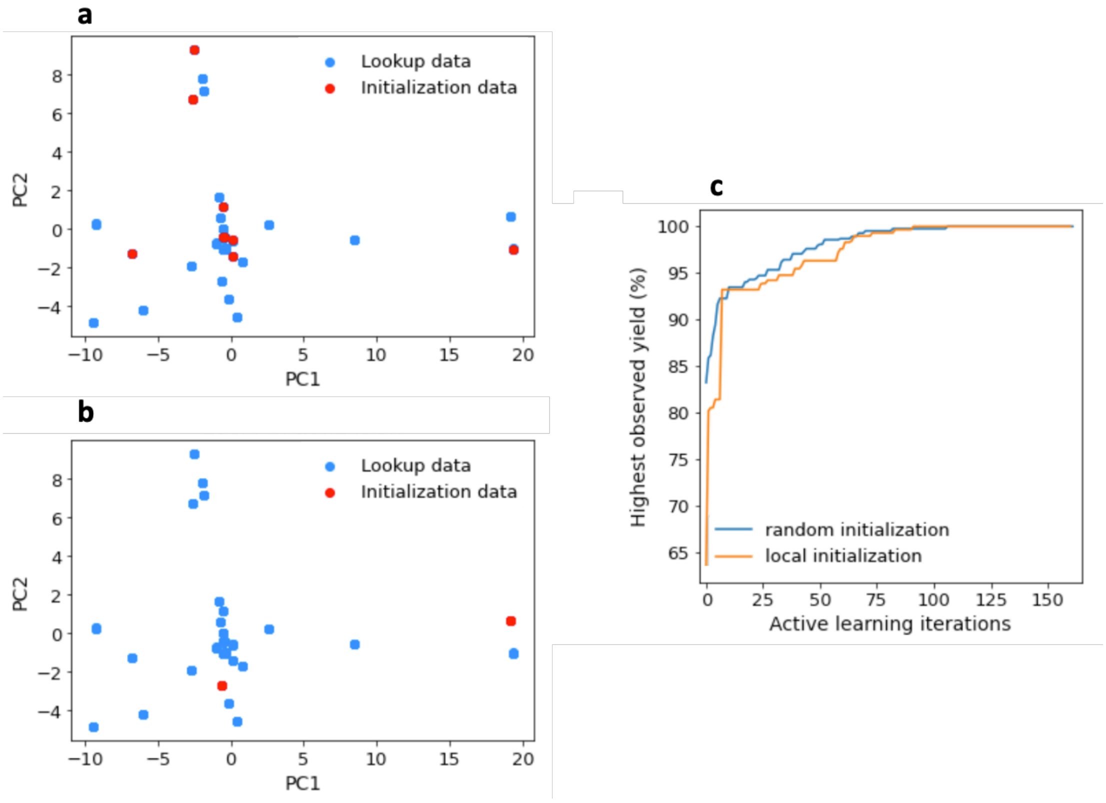


*Figure S19 Random versus local initialization (a) Dimensionality reduced plot of the training and test data within random initialization (b) Dimensionality reduced plot of the training and test data within local initialization (c) Learning curves of the random versus the local initialization*

## The Impact of Initialization: Dataset Size vs. Complexity of Parameterization

We conducted a comparison to understand the performance variation between complexity of the parametrization and the size of the initial dataset. In terms of size of the initialization dataset, 10, 15 and 20 datapoints were chosen along with OHE, Morgan 2 fingerprints and hybrid full feature representation. Figure S20 illustrates all the learning curves of the conducted experiments – in the main manuscript we restricted the plot to 4 trajectories for simplification. The trend of increased performance when using a larger number initialization datapoints using OHE as opposed to a smaller but more complex parameterized initialization dataset could be observed again.

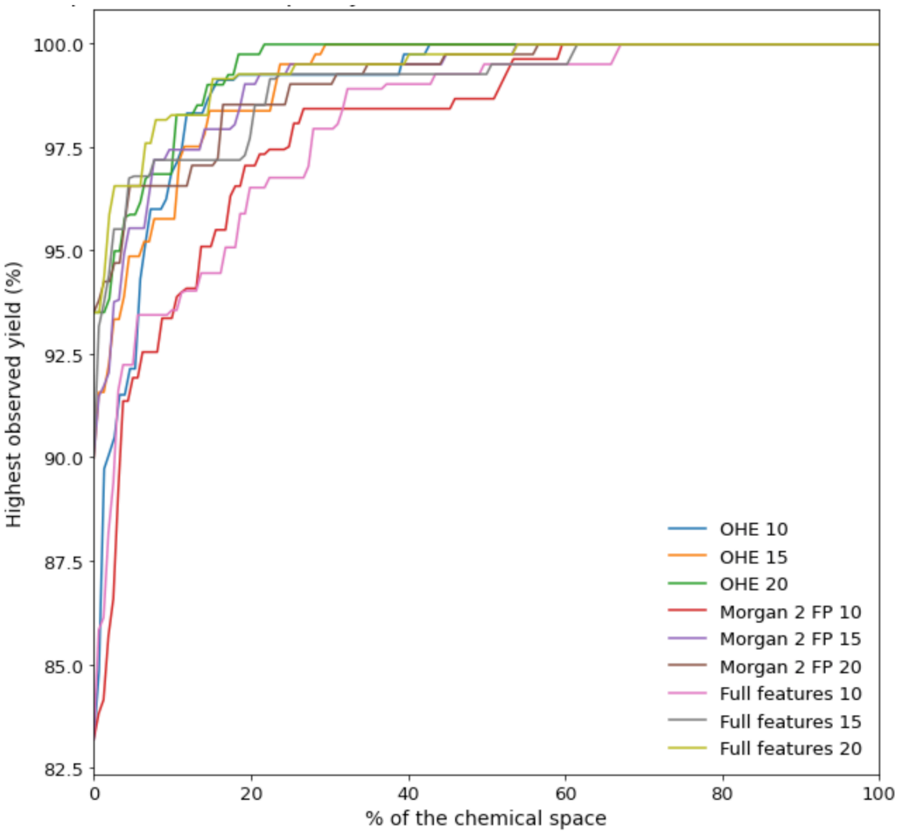


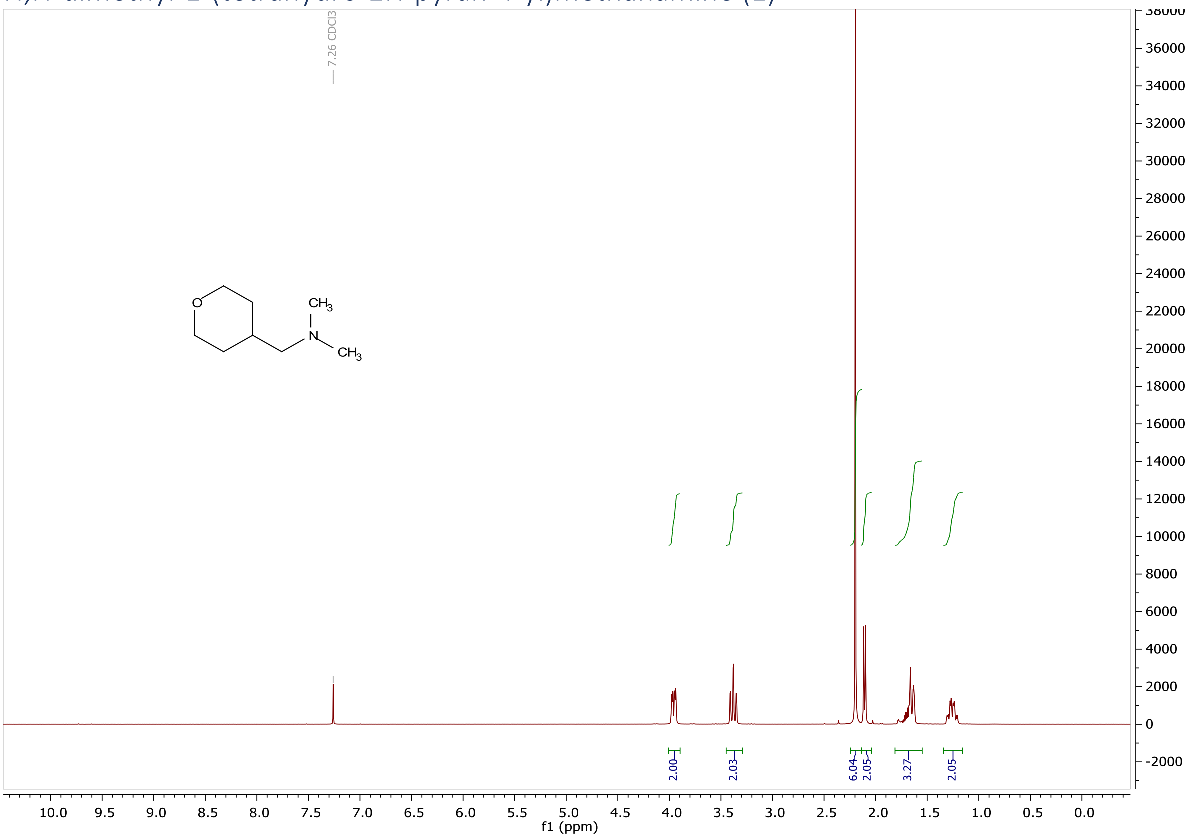
*Figure S20 Evaluation of different initialization strategies for the active learning - variation of chemical representation and size of the initialization dataset*

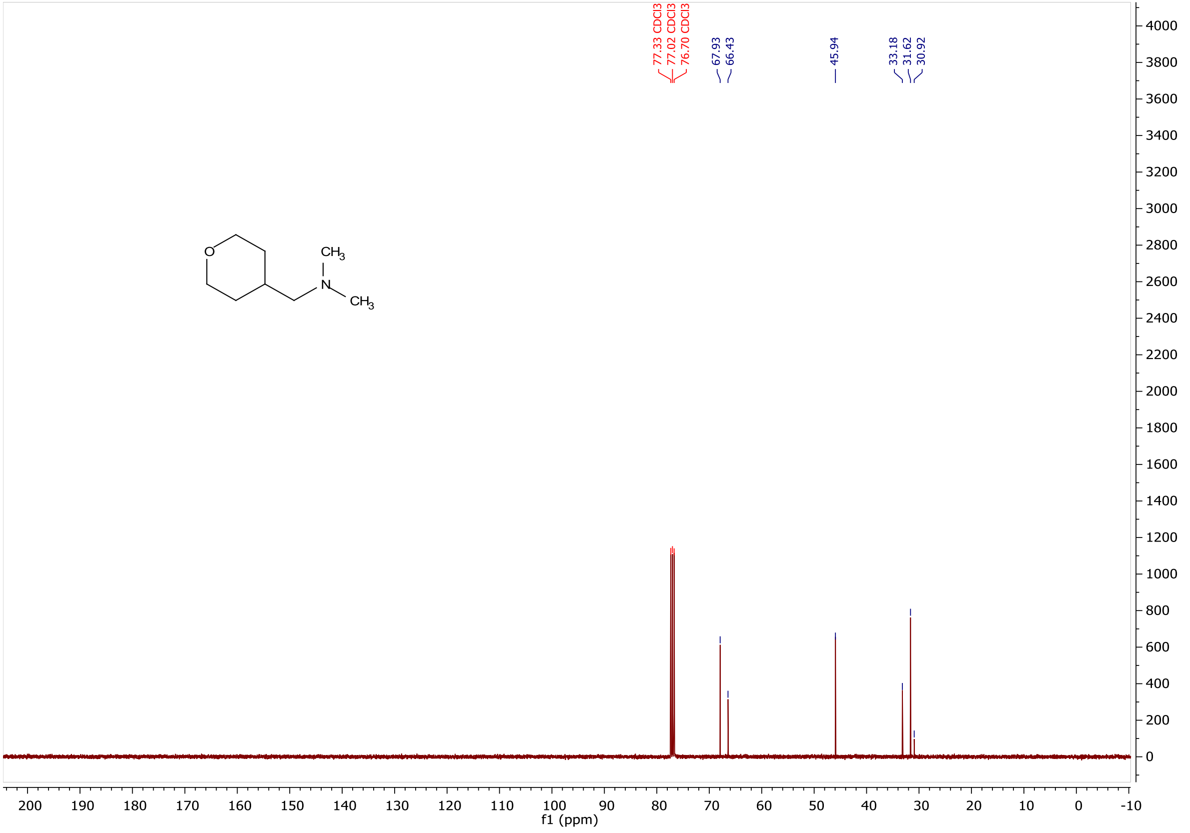
## References

1. D. Y. Ong, Z. Yen, A. Yoshii, J. Revillo Imbernon, R. Takita and S. Chiba, *Angew. Chem. Int. Ed.*, 2019, **58**, 4992-4997.
2. A. V. Brethomé, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313-2323.
3. L. Falivene, Z. Cao, A. Petta, L. Serra, A. Poater, R. Oliva, V. Scarano and L. Cavallo, *Nat. Chem.*, 2019, **11**, 872-879.
4. F. Weinhold, C. R. Landis and E. D. Glendening, *Int. Rev. in Phys. Chem.*, 2016, **35**, 399-440.
5. C. M. Breneman and K. B. Wiberg, *J. of Comp. Chem.*, 1990, **11**, 361-373.
6. G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, *J. of Med. Chem.*, 2014, **57**, 3186-3204.
7. J. Feng, J. L. Lansford, M. A. Katsoulakis and D. G. Vlachos, *Sci. Adv.*, 2020, **6**.
8. J. Jiménez-Luna, F. Grisoni and G. Schneider, *Nat. Mach. Intell.*, 2020, **2**, 573-584.
9. B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich and F. A. Hamprecht, *BMC Bioinform.*, 2009, **10**, 213.

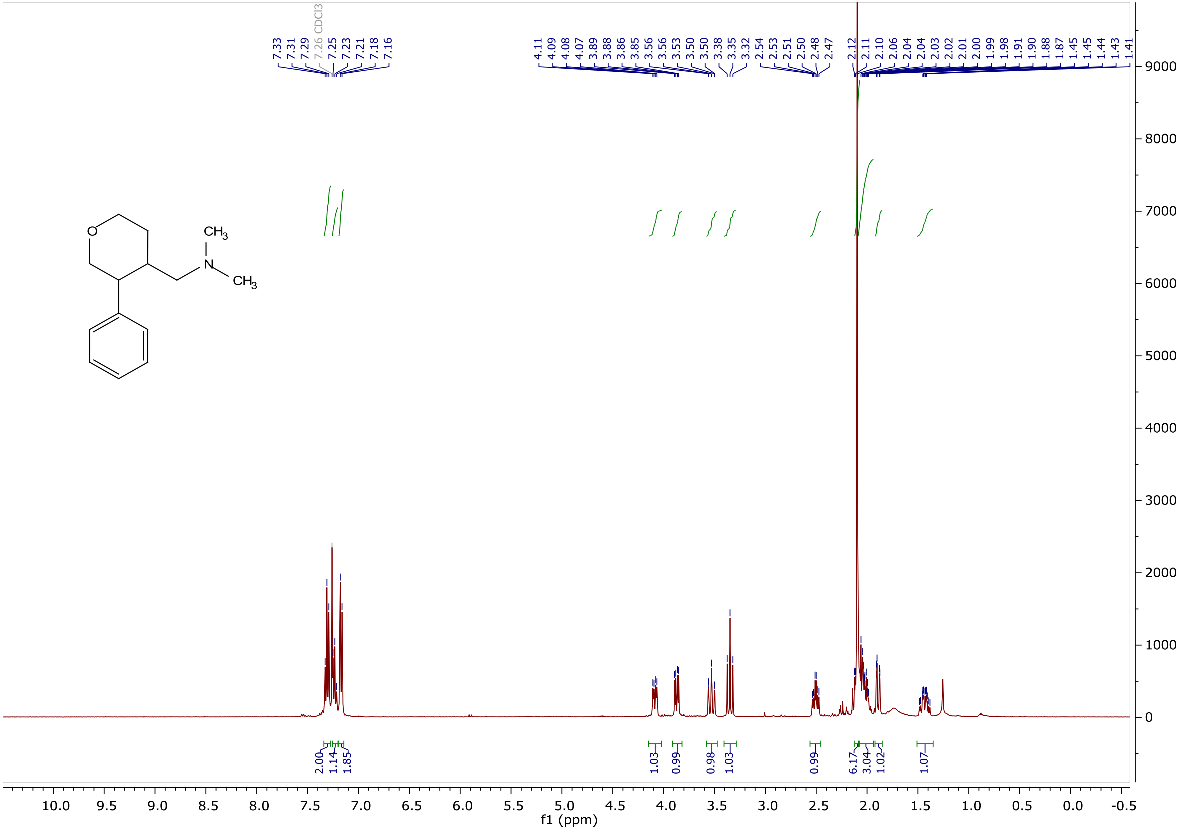
# Appendix

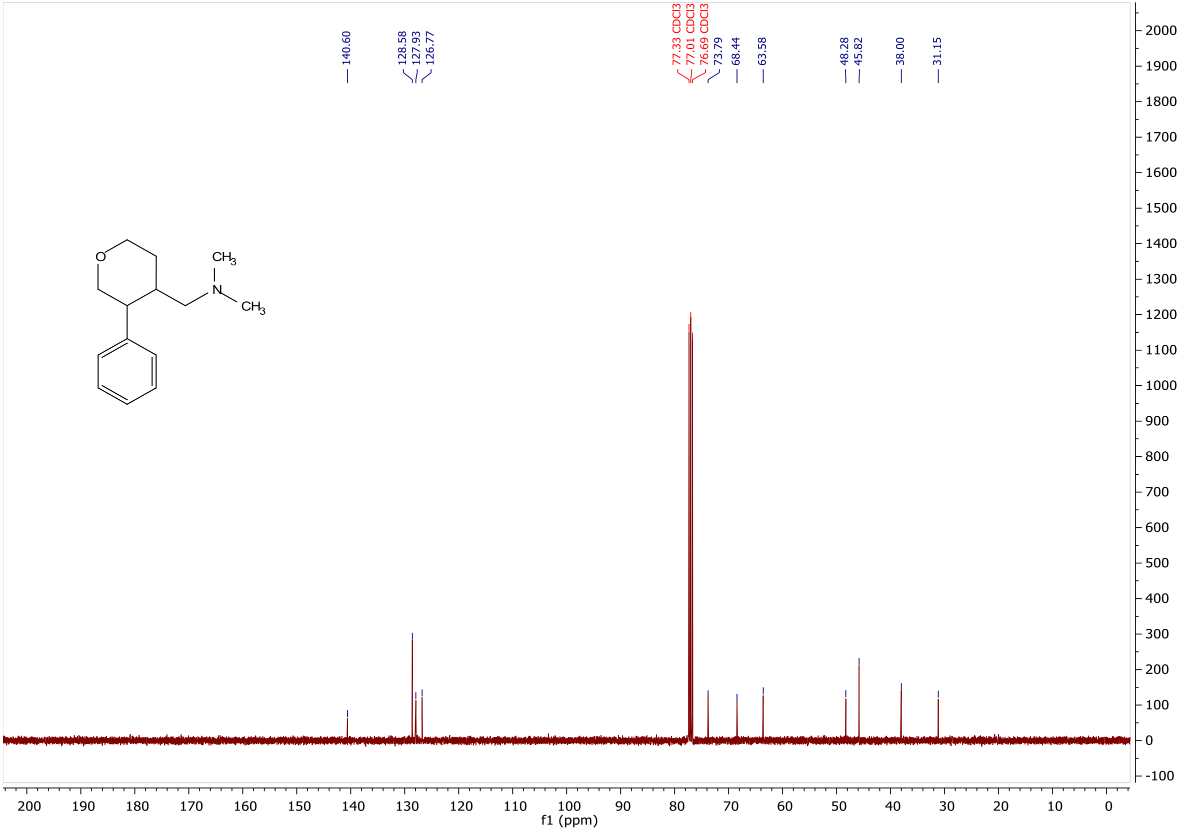
## N,N-dimethyl-1-(tetrahydro-2H-pyran-4-yl)methanamine (1)





## N,N-dimethyl-1-(3-phenyltetrahydro-2H-pyran-4-yl)methanamine (2)


7.26 CDCl3
O
CH3
N
CH3
f1 (ppm)


77.33 CDCl3
77.01 CDCl3
76.69 CDCl3
O
CH3
N
CH3
f1 (ppm)